

Second Edition

 **WILEY**

THE HISTORY OF MATHEMATICS

A Brief Course



WWW.
LINK AVAILABLE

Roger Cooke

The History of Mathematics

The History of Mathematics

A Brief Course

Second Edition

Roger Cooke
University of Vermont



WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2005 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-646-8600, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data:

Cooke, Roger, 1942

The history of mathematics: a brief course/Roger Cooke -- 2nd ed.
p.cm.

Includes bibliographical references and indexes.

ISBN 978-0-471-44459-6

1. Mathematics--History. I. Title

QA 21.C649 2005

510'.9--dc22

2004042299

Contents

Preface	xv
Part 1. The World of Mathematics and the Mathematics of the World	1
Chapter 1. The Origin and Prehistory of Mathematics	3
1. Numbers	4
1.1. Animals' use of numbers	5
1.2. Young children's use of numbers	5
1.3. Archaeological evidence of counting	6
2. Continuous magnitudes	6
2.1. Perception of shape by animals	7
2.2. Children's concepts of space	8
2.3. Geometry in arts and crafts	9
3. Symbols	9
4. Mathematical inference	12
4.1. Visual reasoning	12
4.2. Chance and probability	13
Questions and problems	14
Chapter 2. Mathematical Cultures I	19
1. The motives for creating mathematics	19
1.1. Pure versus applied mathematics	19
2. India	21
2.1. The <i>Sulva Sutras</i>	22
2.2. Buddhist and Jaina mathematics	23
2.3. The Bakshali Manuscript	23
2.4. The <i>siddhantas</i>	23
2.5. Aryabhata I	24
2.6. Brahmagupta	25
2.7. Bhaskara II	25
2.8. Muslim India	26
2.9. Indian mathematics in the colonial period and after	26
3. China	27
3.1. Works and authors	29
3.2. China's encounter with Western mathematics	32
4. Ancient Egypt	34
5. Mesopotamia	35
6. The Maya	37
6.1. The Dresden Codex	37

Questions and problems	38
Chapter 3. Mathematical Cultures II	41
1. Greek and Roman mathematics	41
1.1. Sources	42
1.2. General features of Greek mathematics	43
1.3. Works and authors	45
2. Japan	50
2.1. Chinese influence and calculating devices	51
2.2. Japanese mathematicians and their works	51
3. The Muslims	54
3.1. Islamic science in general	54
3.2. Some Muslim mathematicians and their works	56
4. Europe	58
4.1. Monasteries, schools, and universities	58
4.2. The high Middle Ages	59
4.3. Authors and works	59
5. North America	62
5.1. The United States and Canada before 1867	63
5.2. The Canadian Federation and post Civil War United States	66
5.3. Mexico	69
6. Australia and New Zealand	70
6.1. Colonial mathematics	70
7. The modern era	72
7.1. Educational institutions	72
7.2. Mathematical societies	73
7.3. Journals	73
Questions and problems	73
Chapter 4. Women Mathematicians	75
1. Individual achievements and obstacles to achievement	76
1.1. Obstacles to mathematical careers for women	76
2. Ancient women mathematicians	80
3. Modern European women	81
3.1. Continental mathematicians	82
3.2. Nineteenth-century British women	85
3.3. Four modern pioneers	88
4. American women	100
5. The situation today	104
Questions and problems	105
Part 2. Numbers	109
Chapter 5. Counting	111
1. Number words	111
2. Bases for counting	113
2.1. Decimal systems	113
2.2. Nondecimal systems	114
3. Counting around the world	116
3.1. Egypt	116

3.2. Mesopotamia	116
3.3. India	118
3.4. China	118
3.5. Greece and Rome	119
3.6. The Maya	121
4. What was counted?	122
4.1. Calendars	122
4.2. Weeks	125
Questions and problems	127
Chapter 6. Calculation	129
1. Egypt	129
1.1. Multiplication and division	130
1.2. "Parts"	131
1.3. Practical problems	134
2. China	135
2.1. Fractions and roots	136
2.2. The <i>Jiu Zhang Suanshu</i>	138
3. India	139
4. Mesopotamia	140
5. The ancient Greeks	142
6. The Islamic world	143
7. Europe	143
8. The value of calculation	145
9. Mechanical methods of computation	146
9.1. Software: prosthaphæresis and logarithms	146
9.2. Hardware: slide rules and calculating machines	149
9.3. The effects of computing power	153
Questions and problems	154
Chapter 7. Ancient Number Theory	159
1. Plimpton 322	159
2. Ancient Greek number theory	164
2.1. The <i>Arithmetica</i> of Nicomachus	165
2.2. Euclid's number theory	168
2.3. The <i>Arithmetica</i> of Diophantus	170
3. China	172
4. India	175
4.1. Varahamihira's mystical square	175
4.2. Aryabhata I	175
4.3. Brahmagupta	175
4.4. Bhaskara II	178
5. The Muslims	179
6. Japan	180
7. Medieval Europe	181
Questions and problems	182
Chapter 8. Numbers and Number Theory in Modern Mathematics	187
1. Modern number theory	187
1.1. Fermat	187

1.2. Euler	188
1.3. Lagrange	190
1.4. Legendre	191
1.5. Gauss	192
1.6. Dirichlet	193
1.7. Riemann	194
1.8. Fermat's last theorem	195
1.9. The prime number theorem	196
2. Number systems	197
2.1. Negative numbers and zero	197
2.2. Irrational and imaginary numbers	199
2.3. Imaginary and complex numbers	206
2.4. Infinite numbers	209
3. Combinatorics	210
3.1. Summation rules	210
Questions and problems	217
Part 3. Color Plates	221
Part 4. Space	231
Chapter 9. Measurement	233
1. Egypt	234
1.1. Areas	235
1.2. Volumes	239
2. Mesopotamia	241
2.1. The Pythagorean theorem	242
2.2. Plane figures	243
2.3. Volumes	244
3. China	244
3.1. The <i>Zhou Bi Suan Jing</i>	244
3.2. The <i>Jiu Zhang Suanshu</i>	247
3.3. The <i>Sun Zi Suan Jing</i>	248
3.4. Liu Hui	249
3.5. Zu Chongzhi	250
4. Japan	252
4.1. The challenge problems	252
4.2. Beginnings of the calculus in Japan	253
5. India	257
5.1. Aryabhata I	257
5.2. Brahmagupta	262
Questions and problems	264
Chapter 10. Euclidean Geometry	269
1. The earliest Greek geometry	269
1.1. Thales	270
1.2. Pythagoras and the Pythagoreans	271
1.3. Pythagorean geometry	272
1.4. Challenges to Pythagoreanism: unsolved problems	274

1.5. Challenges to Pythagoreanism: the paradoxes of Zeno of Elea	283
1.6. Challenges to Pythagoreanism: incommensurables	284
1.7. The influence of Plato	285
1.8. Eudoxan geometry	287
1.9. Aristotle	293
2. Euclid	296
2.1. The <i>Elements</i>	296
2.2. The <i>Data</i>	299
3. Archimedes	299
3.1. The area of a sphere	301
3.2. The <i>Method</i>	302
4. Apollonius	304
4.1. History of the <i>Conics</i>	305
4.2. Contents of the <i>Conics</i>	305
4.3. Apollonius' definition of the conic sections	306
4.4. Foci and the three- and four-line locus	308
Questions and problems	310
 Chapter 11. Post-Euclidean Geometry	 317
1. Hellenistic geometry	318
1.1. Zenodorus	318
1.2. The parallel postulate	319
1.3. Heron	320
1.4. Pappus	322
2. Roman geometry	325
2.1. Roman civil engineering	327
3. Medieval geometry	328
3.1. Late Medieval and Renaissance geometry	330
4. Geometry in the Muslim world	332
4.1. The parallel postulate	333
4.2. Thabit ibn-Qurra	333
4.3. Al-Kuhi	335
4.4. Al-Haytham	335
4.5. Omar Khayyam	336
4.6. Nasir al-Din al-Tusi	337
5. Non-Euclidean geometry	338
5.1. Girolamo Saccheri	339
5.2. Lambert and Legendre	341
5.3. Gauss	342
5.4. Lobachevskii and János Bolyai	343
5.5. The reception of non-Euclidean geometry	346
5.6. Foundations of geometry	348
6. Questions and problems	348
 Chapter 12. Modern Geometries	 351
1. Analytic and algebraic geometry	351
1.1. Fermat	351
1.2. Descartes	352
1.3. Newton's classification of curves	355

1.4. Algebraic geometry	355
2. Projective and descriptive geometry	356
2.1. Projective properties	356
2.2. The Renaissance artists	357
2.3. Girard Desargues	360
2.4. Blaise Pascal	364
2.5. Newton's degree-preserving mappings	364
2.6. Charles Brianchon	365
2.7. Monge and his school	366
2.8. Jacob Steiner	367
2.9. August Ferdinand Möbius	368
2.10. Julius Plücker	369
2.11. Arthur Cayley	371
3. Differential geometry	371
3.1. Huygens	371
3.2. Newton	373
3.3. Leibniz	374
3.4. The eighteenth century	375
3.5. Gauss	376
3.6. The French and British geometers	379
3.7. Riemann	380
3.8. The Italian geometers	383
4. Topology	385
4.1. Early combinatorial topology	386
4.2. Riemann	386
4.3. Möbius	388
4.4. Poincaré's <i>Analysis situs</i>	388
4.5. Point-set topology	390
Questions and problems	393
 Part 5. Algebra	 397
Chapter 13. Problems Leading to Algebra	399
1. Egypt	399
2. Mesopotamia	401
2.1. Linear and quadratic problems	401
2.2. Higher-degree problems	403
3. India	404
3.1. Jaina algebra	404
3.2. The Bakshali Manuscript	404
4. China	405
4.1. The <i>Jiu Zhang Suanshu</i>	405
4.2. The <i>Suanshu Shu</i>	405
4.3. The <i>Sun Zi Suan Jing</i>	406
4.4. Zhang Qiujiang	406
Questions and problems	407
 Chapter 14. Equations and Algorithms	 409
1. The <i>Arithmetica</i> of Diophantus	409

1.1.	Diophantine equations	410
1.2.	General characteristics of the <i>Arithmetica</i>	410
1.3.	Determinate problems	411
1.4.	The significance of the <i>Arithmetica</i>	412
1.5.	The view of Jacob Klein	412
2.	China	413
2.1.	Linear equations	413
2.2.	Quadratic equations	413
2.3.	Cubic equations	414
2.4.	The numerical solution of equations	415
3.	Japan	417
3.1.	Seki Kōwa	418
4.	Hindu algebra	420
4.1.	Brahmagupta	420
4.2.	Bhaskara II	421
5.	The Muslims	422
5.1.	Al-Khwarizmi	423
5.2.	Abu Kamil	425
5.3.	Omar Khayyam	425
5.4.	Sharaf al-Din al-Muzaffar al-Tusi	426
6.	Europe	427
6.1.	Leonardo of Pisa (Fibonacci)	428
6.2.	Jordanus Nemorarius	429
6.3.	The fourteenth and fifteenth centuries	429
6.4.	Chuquet	430
6.5.	Solution of cubic and quartic equations	431
6.6.	Consolidation	432
	Questions and problems	434
Chapter 15.	Modern Algebra	437
1.	Theory of equations	437
1.1.	Albert Girard	437
1.2.	Tschirnhaus transformations	438
1.3.	Newton, Leibniz, and the Bernoullis	440
1.4.	Euler, d'Alembert, and Lagrange	440
1.5.	Gauss and the fundamental theorem of algebra	443
1.6.	Ruffini	444
1.7.	Cauchy	444
1.8.	Abel	446
1.9.	Galois	447
2.	Algebraic structures	451
2.1.	Fields, rings, and algebras	451
2.2.	Abstract groups	454
2.3.	Number systems	458
	Questions and problems	459
Part 6.	Analysis	461
Chapter 16.	The Calculus	463

1. Prelude to the calculus	463
1.1. Tangent and maximum problems	464
1.2. Lengths, areas, and volumes	465
1.3. The relation between tangents and areas	467
1.4. Infinite series and products	467
2. Newton and Leibniz	468
2.1. Isaac Newton	468
2.2. Gottfried Wilhelm von Leibniz	470
2.3. The disciples of Newton and Leibniz	472
3. Branches and roots of the calculus	475
3.1. Ordinary differential equations	475
3.2. Partial differential equations	477
3.3. Calculus of variations	478
3.4. Foundations of the calculus	483
Questions and problems	487
 Chapter 17. Real and Complex Analysis	 489
1. Complex analysis	489
1.1. Algebraic integrals	490
1.2. Cauchy	493
1.3. Riemann	494
1.4. Weierstrass	495
2. Real analysis	496
2.1. Fourier series, functions, and integrals	496
2.2. Completeness of the real numbers	502
2.3. Uniform convergence and continuity	503
2.4. General integrals and discontinuous functions	503
2.5. The abstract and the concrete	504
2.6. Discontinuity as a positive property	506
Questions and problems	507
 Part 7. Mathematical Inferences	 509
 Chapter 18. Probability and Statistics	 511
1. Probability	511
1.1. Cardano	512
1.2. Fermat and Pascal	513
1.3. Huygens	514
1.4. Leibniz	515
1.5. The <i>Ars Conjectandi</i> of Jakob Bernoulli	515
1.6. De Moivre	517
1.7. Laplace	520
1.8. Legendre	521
1.9. Gauss	521
1.10. Philosophical issues	522
1.11. Large numbers and limit theorems	523
2. Statistics	524
2.1. Quetelet	525
2.2. Statistics in physics	525

2.3. The metaphysics of probability and statistics	527
2.4. Correlations and statistical inference	528
Questions and problems	531
Chapter 19. Logic and Set Theory	535
1. Logic	535
1.1. From algebra to logic	535
1.2. Symbolic calculus	539
1.3. Boole's <i>Mathematical Analysis of Logic</i>	539
1.4. Boole's <i>Laws of Thought</i>	541
1.5. Venn	542
1.6. Jevons	544
2. Set theory	544
2.1. Technical background	544
2.2. Cantor's work on trigonometric series	545
2.3. The reception of set theory	547
2.4. Existence and the axiom of choice	548
2.5. Doubts about set theory	551
3. Philosophies of mathematics	552
3.1. Paradoxes	553
3.2. Formalism	554
3.3. Intuitionism	555
3.4. Mathematical practice	556
Questions and problems	557
Literature	561
Subject Index	581
Name Index	599

Preface

This second edition of *The History of Mathematics: A Brief Course* must begin with a few words of explanation to all users of the first edition. The present volume constitutes such an extensive rewriting of the original that it amounts to a considerable stretch in the meaning of the phrase *second edition*. Although parts of the first edition have been retained, I have completely changed the order of presentation of the material. A comparison of the two tables of contents will reveal the difference at a glance: In the first edition each chapter was devoted to a single culture or period within a single culture and subdivided by mathematical topics. In this second edition, after a general survey of mathematics and mathematical practice in Part 1, the primary division is by subject matter: numbers, geometry, algebra, analysis, mathematical inference.

For reasons that mathematics can illustrate very well, writing the history of mathematics is a nearly impossible task. To get a proper orientation for any particular event in mathematical history, it is necessary to take account of three independent “coordinates”: the time, the mathematical subject, and the culture. To thread a narrative that is to be read linearly through this three-dimensional array of events is like drawing one of Peano’s space-filling curves. Some points on the curve are infinitely distant from one another, and the curve must pass through some points many times. From the point of view of a reader whose time is valuable, these features constitute a glaring defect. The problem is an old one, well expressed eighty years ago by Felix Klein, in Chapter 6 of his *Lectures on the Development of Mathematics in the Nineteenth Century*:

I have now mentioned a large number of more or less famous names, all closely connected with Riemann. They can become more than a mere list only if we look into the literature associated with the names, or rather, with those who bear the names. One must learn how to grasp the main lines of the many connections in our science out of the enormous available mass of printed matter without getting lost in the time-consuming discussion of every detail, but also without falling into superficiality and dilettantism.

Klein writes as if it were possible to achieve this laudable goal, but then his book was by intention only a collection of essays, not a complete history. Even so, he used more pages to tell the story of one century of European mathematics than a modern writer has available for the history of all of mathematics. For a writer who hates to leave any threads dangling the necessary sacrifices are very painful. My basic principle remains the same as in the first edition: not to give a mere list of names and results described in general terms, but to show the reader what important results were achieved and in what context. Even if unlimited pages

were available, time is an important consideration for authors as well as readers. To switch metaphors, there were so many times during the writing when tempting digressions arose which I could not resist pursuing, that I suspected that I might be traversing the boundary of a fractal snowflake or creating the real-life example of Zeno's dichotomy. Corrections and supplementary material relating to this book can be found at my website at the University of Vermont. The url is:

<http://www.cem.uvm.edu/~cooke/history/seconded.html>

Fortunately, significant mathematical events are discrete, not continuous, so that a better analogy for a history of mathematics comes from thermodynamics. If the state of mathematics at any given time is a system, its atoms are mathematical problems and propositions, grouped into molecules of theory. As they evolve, these molecules sometimes collide and react chemically, as happened with geometry and algebra in the seventeenth century. The resulting development of the mathematical system resembles a Brownian motion; and while it is not trivial to describe a Brownian motion in detail, it is easier than drawing a space-filling curve.

Now let me speak more literally about what I have tried to do in the present book. As mentioned above, Part 1 is devoted to a broad survey of the world of mathematics. Each of the six subsequent parts, except Part 3, where the color plates are housed, concentrates on a particular aspect of mathematics (arithmetic, geometry, algebra, analysis, and mathematical inference) and discusses its development in different cultures over time. I had two reasons for reorganizing the material in this way.

First, I am convinced that students will remember better what they learn if they can focus on a single area of mathematics, comparing what was done in this area by different cultures, rather than studying the arithmetic, geometry, and algebra of each culture by turns. Second, although reviewers were for the most part kind, I was dissatisfied with the first edition, feeling that the organization of the book along cultural lines had caused me to omit many good topics, especially biographical material, and sources that really ought to have been included. The present edition aims to correct these omissions, along with a number of mistakes that I have noticed or others have pointed out. I hope that the new arrangement of material will make it possible to pursue the development of a single area of mathematics to whatever level the instructor wishes, then turn to another area and do the same. A one-semester course in mostly elementary mathematics from many cultures could be constructed from Chapters 1–7, 9–11, and 13–14. After that, one could use any remaining time to help the students write term papers (which I highly recommend) or go on to read other chapters in the book. I would also point out that, except for Chapters 8–12, and 15–19, the chapters, and even the sections within the chapters, can be read independently of one another. For a segment on traditional Chinese mathematics, for example, students could be assigned Section 3 of Chapter 2, Subsection 3.4 of Chapter 5, Section 2 of Chapter 6, Section 3 of Chapter 7, Section 3 of Chapter 9, Section 4 of Chapter 13, and Section 2 of Chapter 14.

Because of limitations of time and space, the present book will show the reader only a few of the major moments in the history of mathematics, omitting many talented mathematicians and important results. This restriction to the important moments makes it impossible to do full justice to what Grattan-Guinness has stated as the question the historian should answer: What happened in the past? We are reconstructing an evolutionary process, but the "fossil record" presented in

any general history of mathematics will have many missing links. Unavoidably, history gets distorted in this process. New results appear more innovative than they actually are. To take just one example (not discussed elsewhere in this book), it was a very clever idea of Hermann Weyl to trivialize the proof of Kronecker's theorem that the fractional parts of the multiples of an irrational number are uniformly distributed in the unit interval; Weyl made this result a theorem about discrete and continuous averages of integrable periodic functions. One would expect that in an evolutionary process, there might be an intermediate step—someone who realized that these fractional parts are dense but not necessarily that they are uniformly distributed. And indeed there was: Nicole d'Oresme, 500 years before Kronecker. There are hundreds of results in mathematics with names on them, in many cases incorrectly attributed, and in many more cited in a much more polished form than the discoverer ever imagined. History ought to correct this misimpression, but a general history has only a limited ability to do so.

The other question mentioned by Grattan-Guinness—How did things come to be the way they are?—is often held up in history books as the main justification for requiring students to study political and social history.¹ That job is somewhat easier to do in a general textbook, and I hope the reader will be pleased to learn how some of the current parts of the curriculum arose.

I would like to note here three small technical points about the second edition.

Citations. In the first edition I placed a set of endnotes in each chapter telling the sources from which I had derived the material of that chapter. In the present edition I have adopted the more scholarly practice of including a bibliography organized by author and date. In the text itself, I include citations at the points where they are used. Thus, the first edition of this book would be cited as (Cooke, 1997). Although I dislike the interruption of the narrative that this practice entails, I do find it convenient when reading the works of others to be able to note the source of a topic that I think merits further study without having to search for the citation. On balance, I think the advantage of citing a source on the spot outweighs the disadvantage of having to block out parenthetical material in order to read the narrative.

Translations. Unless another source is cited, all translations from foreign languages are my own. The reader may find smoother translations in most cases. To bring out significant concepts, especially in quotations from ancient Greek, I have made translations that are more literal than the standard ones. Since I don't know Sanskrit, Arabic, or Chinese, the translations from those languages are not mine; the source should be clear from the surrounding text.

Cover. Wiley has done me the great favor of producing a cover design in four colors rather than the usual two. That consideration made it possible to use a picture that I took at a quilt exposition at Norwich University (Northfield, Vermont) in 2003. The design bears the title "A Number Called Phi," and its creator, Mary Knapp of Watertown, New York, incorporated many interesting mathematical connections through the geometric and floral shapes it contains. I am grateful for her permission to use it as the cover of this second edition.

¹ In a lecture at the University of Vermont in September 2003 Grattan-Guinness gave the name *heritage* to the attempt to answer this question. *Heritage* is a perfectly respectable topic to write on, but the distinction between history and heritage is worth keeping in mind. See his article on this distinction (Grattan-Guinness, 2004).

Acknowledgements. I am grateful to the editors at Wiley, Steve Quigley and especially Susanne Steitz, for keeping in touch throughout the long period of preparation for this book. I would also like to thank the copy editor, Barbara Zeiders, who made so many improvements to the text that I could not begin to list them all. I also ask Barbara to forgive my obstinacy on certain issues involving commas, my stubborn conviction that *independently of* is correct usage, and my constitutional inability to make all possessives end in 's. I cannot bring myself to write *Archimedes's* or *Descartes's*; and if we are going to allow some exceptions for words ending in a sibilant, as we must, I prefer—in defiance of the *Chicago Manual of Style*—to use an unadorned apostrophe for the possessive of *all* words ending in *s*, *z*, or *x*.

The diagram of Florence Nightingale's statistics on the Crimean War (Plate 5) is in the public domain; I wish to thank *Cabinet* magazine for providing the electronic file for this plate.

Many of the literature references in the chapters that follow were given to me by the wonderful group of mathematicians and historians on the Historia Mathematica e-mail list. It seemed that, no matter how obscure the topic on which I needed information, there was someone on the list who knew something about it. To Julio Gonzalez Cabillon, who maintains the list as a service to the community, I am deeply grateful.

Roger Cooke

January 2005

The History of Mathematics: A Brief Course, Second Edition

by Roger Cooke

Copyright © 2005 John Wiley & Sons, Inc.

Part 1

The World of Mathematics and the Mathematics of the World

This first part of our history is concerned with the “front end” of mathematics (to use an image from computer algebra)—its relation to the physical world and human society. It contains some general considerations about mathematics, what it consists of, how it may have arisen, and how it has developed in various cultures around the world. Because of the large number of cultures that exist, a considerable paring down of the available material is necessary. We are forced to choose a few sample cultures to represent the whole, and we choose those that have the best-recorded mathematical history. The general topics studied in this part involve philosophical and social questions, which are themselves specialized subjects of study, to which a large amount of scholarly literature has been devoted. Our approach here is the naive commonsense approach of an author who is not a specialist in either philosophy or sociology. Since present-day governments have to formulate *policies* relating to mathematics and science, it is important that such questions not be left to specialists. The rest of us, as citizens of a republic, should read as much as time permits of what the specialists have to say and make up our own minds when it comes time to judge the effects of a policy.

This section consists of four chapters. In Chapter 1 we consider the nature and prehistory of mathematics. In this area we are dependent on archaeologists and anthropologists for the comparatively small amount of historical information available. We ask such questions as the following: “What is the subject matter of mathematics?” “Is new mathematics created to solve practical problems, or is it an expression of free human imagination, or some of each?” “How are mathematical concepts related to the physical world?”

Chapter 2 begins a broad survey of mathematics around the world. This chapter is subdivided according to a selection of cultures in which mathematics has arisen as an indigenous creation, in which borrowings from other cultures do not play a prominent role. For each culture we give a summary of the development of mathematics in that culture, naming the most prominent mathematicians and their works. Besides introducing the major works and their authors, an important goal of this chapter is to explore the question, “Why were these works written?” We quote the authors themselves as often as possible to bring out their motives. Chapters 2 and 3 are intended as background for the topic-based presentation that follows beginning with Chapter 5.

In Chapter 3 we continue the survey with a discussion of mathematical cultures that began on the basis of knowledge and techniques that had been created elsewhere. The contributions made by these cultures are found in the extensions, modifications, and innovations—some very ingenious—added to the inherited materials. In dividing the material over two chapters we run the risk of seeming to minimize the creations of these later cultures. Creativity is involved in mathematical innovations at every stage, from earliest to latest. The reason for having two chapters instead of one is simply that there is too much material for one chapter.

Chapter 4 is devoted to the special topic of women mathematicians. Although the *subject* of mathematics is gender neutral in the sense that no one could determine the gender of the author of a mathematical paper from an examination of the mathematical arguments given, the *profession* of mathematics has not been and is not yet gender neutral. There are obvious institutional and cultural explanations for this fact; but when an area of human endeavor has been polarized by gender, as mathematics has been, that feature is an important part of its history and deserves special attention.

CHAPTER 1

The Origin and Prehistory of Mathematics

In this chapter we have two purposes: first, to consider what mathematics is, and second, to examine some examples of *protomathematics*, the kinds of mathematical thinking that people naturally engage in while going about the practical business of daily life. This agenda assumes that there is a mode of thought called *mathematics* that is intrinsic to human nature and common to different cultures. The simplest assumption is that counting and common shapes such as squares and circles have the same meaning to everyone. To fit our subject into the space of a book of moderate length, we partition mathematical modes of thought into four categories:

Number. The concept of number is almost always the first thing that comes to mind when mathematics is mentioned. From the simplest finger counting by pre-school children to the recent sophisticated proof of Fermat's last theorem (a theorem at last!), numbers are a fundamental component of the world of mathematics.

Space. It can be argued that space is not so much a "thing" as a convenient way of organizing physical objects in the mind. Awareness of spatial relations appears to be innate in human beings and animals, which must have an instinctive understanding of space and time in order to move purposefully. When people began to intellectualize this intuitive knowledge, one of the first efforts to organize it involved reducing geometry to arithmetic. Units of length, area, volume, weight, and time were chosen, and *measurement* of these continuous quantities was reduced to *counting* these imaginatively constructed units. In all practical contexts measurement becomes counting in exactly this way. But in pure thought there is a distinction between what is *infinitely divisible* and what is *atomic* (from the Greek word meaning *indivisible*). Over the 2500 years that have elapsed since the time of Pythagoras this collision between the discrete modes of thought expressed in arithmetic and the intuitive concept of continuity expressed in geometry has led to puzzles, and the solution of those puzzles has influenced the development of geometry and analysis.

Symbols. Although early mathematics was discussed in ordinary prose, sometimes accompanied by sketches, its usefulness in science and society increased greatly when symbols were introduced to mimic the mental operations performed in solving problems. Symbols for numbers are almost the only *ideograms* that exist in languages written with a phonetic alphabet. In contrast to ordinary words, for example, the symbol 8 stands for an idea that is the same to a person in Japan, who reads it as *hachi*, a person in Italy who reads it as *otto*, and a person in Russia, who reads it as *vosem'*. The introduction of symbols such as $+$ and $=$ to stand for the common operations and relations of mathematics has led to both the clarity that mathematics has for its initiates and the obscurity it suffers from in the eyes of the nonmathematical. Although it is primarily in studying algebra that we become aware of the use of symbolism, symbols are used in other areas, and algebra,

considered as the study of processes inverse to those of arithmetic, was originally studied without symbols.

Symbol-making has been a habit of human beings for thousands of years. The wall paintings on caves in France and Spain are an early example, even though one might be inclined to think of them as pictures rather than symbols. It is difficult to draw a line between a painting such as the *Mona Lisa*, an animé representation of a human being, and the ideogram for a person used in languages whose written form is derived from Chinese. The last certainly *is* a symbol, the first two usually are not thought of that way. Phonetic alphabets, which establish a symbolic, visual representation of sounds, are another early example of symbol-making. A similar spectrum presents itself in the many ways in which human beings convey instructions to one another, the purest being a computer program. Very often, people who think they are not mathematical are quite good at reading abstractly written instructions such as music, blueprints, road maps, assembly instructions for furniture, and clothing patterns. All these symbolic representations exploit a basic human ability to make correspondences and understand analogies.

Inference. Mathematical reasoning was at first numerical or geometric, involving either counting something or “seeing” certain relations in geometric figures. The finer points of logical reasoning, rhetoric, and the like, belonged to other areas of study. In particular, philosophers had charge of such notions as cause, implication, necessity, chance, and probability. But with the Pythagoreans, verbal reasoning came to permeate geometry and arithmetic, supplementing the visual and numerical arguments. There was eventually a countercurrent, as mathematics began to influence logic and probability arguments, eventually producing specialized mathematical subjects: mathematical logic, set theory, probability, and statistics. Much of this development took place in the nineteenth century and is due to mathematicians with a strong interest and background in philosophy. Philosophers continue to speculate on the meaning of all of these subjects, but the parts of them that belong to mathematics are as solidly grounded (apart from their applications) as any other mathematics.

We shall now elaborate on the origin of each of these components. Since these origins are in some cases far in the past, our knowledge of them is indirect, uncertain, and incomplete. A more detailed study of all these areas begins in Part 2. The present chapter is confined to generalities and conjectures as to the state of mathematical knowledge preceding these records.

1. Numbers

Counting objects that are distinct but similar in appearance, such as coins, goats, and full moons, is a universal human activity that must have begun to occur as soon as people had language to express numbers. In fact, it is impossible to imagine that numbers could have arisen without this kind of counting. Several closely related threads can be distinguished in the fabric of elementary arithmetic. First, there is a distinction that we now make between cardinal and ordinal numbers. We think of cardinal numbers as applying to *sets* of things—the word *sets* is meant here in its ordinary sense, not the specialized meaning it has in mathematical set theory—and ordinal numbers as applying to the individual elements of a set by virtue of an ordering imposed on the set. Thus, the cardinal number of the set $\{a, b, c, d, e, f, g\}$ is 7, and *e* is the fifth element of this set by virtue of the standard

alphabetical ordering. These two notions are not so independent as they may appear in this illustration, however. Except for very small sets, whose cardinality can be perceived immediately, the cardinality of a set is usually determined by *counting*, that is, arranging its elements linearly as first, second, third, and so on, even though it may be the corresponding cardinal numbers—one, two, three, and so on—that one says aloud when doing the counting.

A second thread closely intertwined with counting involves the elementary operations of arithmetic. The commonest actions that are carried out with any collection of things are taking objects out of it and putting new objects into it. These actions, as everyone recognizes immediately, correspond to the elementary operations of subtraction and addition. The etymology of these words shows their origin, *subtraction* having the meaning of *pulling out* (literally pulling up or under) and *addition* meaning *giving to*. All of the earliest mathematical documents use addition and subtraction without explanation. The more complicated operations of multiplication and division may have arisen from comparison of two collections of different sizes (counting the number of times that one collection fits into another, or copying a collection a fixed number of times and counting the result), or perhaps as a shortened way of performing addition or subtraction. It is impossible to know much for certain, since most of the early documents also assume that multiplication of small integers is understood without explanation. A notable exception occurs in certain ancient Egyptian documents, where computations that would now be performed using multiplication or division are reduced to repeated doubling, and the details of the computation are shown.

1.1. Animals' use of numbers. Counting is so useful that it has been observed not only in very young children, but also in animals and birds. It is not clear just how high animals and birds can count, but they certainly have the ability to distinguish not merely patterns, but actual numbers. The counting abilities of birds were studied in a series of experiments conducted in the 1930s and 1940s by O. Koehler (1889–1974) at the University of Freiburg. Koehler (1937) kept the trainer isolated from the bird. In the final tests, after the birds had been trained, the birds were filmed automatically, with no human beings present. Koehler found that parrots and ravens could learn to compare the number of dots, up to 6, on the lid of a hopper with a “key” pattern in order to determine which hopper contained food. They could make the comparison no matter how the dots were arranged, thereby demonstrating an ability to take account of the *number* of dots rather than the *pattern*.

1.2. Young children's use of numbers. Preschool children also learn to count and use small numbers. The results of many studies have been summarized by Karen Fuson (1988). A few of the results from observation of children at play and at lessons were as follows:

1. A group of nine children from 21 to 45 months was found to have used the word *two* 158 times, the word *three* 47 times, the word *four* 18 times, and the word *five* 4 times.
2. The children seldom had to count “one–two” in order to use the word *two* correctly; for the word *three* counting was necessary about half the time; for the word *four* it was necessary most of the time; for higher numbers it was necessary all the time.

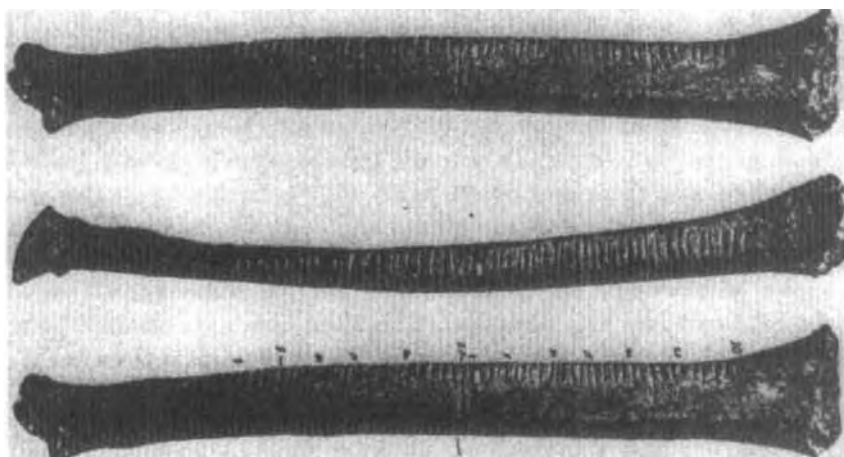
One can thus observe in children the capacity to recognize groups of two or three without performing any conscious numerical process. This observation suggests that these numbers are primitive, while larger numbers are a conscious creation. It also illustrates what was said above about the need for arranging a collection in some linear order so as to be able to find its cardinal number.

1.3. Archaeological evidence of counting. Very ancient animal bones containing notches have been found in Africa and Europe, suggesting that some sort of counting procedure was being carried on at a very early date, although what exactly was being counted remains unknown. One such bone, the radius bone of a wolf, was discovered at Veronice (Czech Republic) in 1937. This bone was marked with two series of notches, grouped by fives, the first series containing five groups and the second six. Its discoverer, Karel Absolon (1887–1960), believed the bone to be about 30,000 years old, although other archaeologists thought it considerably younger. The people who produced this bone were clearly a step above mere survival, since a human portrait carved in ivory was found in the same settlement, along with a variety of sophisticated tools. Because of the grouping by fives, it seems likely that this bone was being used to count something. Even if the groupings are meant to be purely decorative, they point to a use of numbers and counting for a practical or artistic purpose.

Another bone, named after the fishing village of Ishango on the shore of Lake Edward in Zaire where it was discovered in 1960 by the Belgian archaeologist Jean de Heinzelin de Braucourt (1920–1998), is believed to be between 8500 and 11,000 years old. The Ishango Bone, which is now in the Musée d'Histoire Naturelle in Brussels, contains three columns of notches. One column consists of four series of notches containing 11, 21, 19, and 9 notches. Another consists of four series containing 11, 13, 17, and 19 notches. The third consists of eight series containing 3, 6, 4, 8, 10, 5, 5, and 7 notches, with larger gaps between the second and third series and between the fourth and fifth series. These columns present us with a mystery. Why were they put there? What activity was being engaged in by the person who carved them? Conjectures range from abstract experimentation with numbers to keeping score in a game. The bone could have been merely decorative, or it could have been a decorated tool. Whatever its original use, it comes down to the present generation as a reminder that human beings were engaging in abstract thought and creating mathematics a very, very long time ago.

2. Continuous magnitudes

In addition to the ability to count, a second important human faculty is the ability to perceive spatial and temporal relations. These perceptions differ from the discrete objects that elicit counting behavior in that the objects involved are perceived as being divisible into arbitrarily small parts. Given any length, one can always imagine cutting it in half, for example, to get still smaller lengths. In contrast, a penny cut in half does not produce two coins each having a value of one-half cent. Just as human beings are endowed with the ability to reason numerically and understand the concept of equal distribution of money or getting the correct change with a purchase, it appears that we also have an innate ability to reason spatially, for example, to understand that two areas are equal even when they have different shapes, provided that they can be dissected into congruent pieces, or that



The Veronica wolf bone, from the *Illustrated London News*, October 2, 1937.

two vessels of different shape have the same volume if one each holds exactly enough water to fill the other.

One important feature of counting as opposed to measuring—arithmetic as opposed to geometry—is its exactitude. Two sets having the same number of members are numerically *exactly equal*. In contrast, one cannot assert that two sticks, for example, are *exactly* the same length. This difference arises in countless contexts important to human society. Two people may have exactly the same amount of money in the bank, and one can make such an assertion with complete confidence after examining the balance of each of them. But it is only within some limit of error that one could assert that two people are of the same height. The word *exact* would be inappropriate in this context. The notion of absolute equality in relation to continuous objects means *infinite precision* and can be expressed only through the concept of a *real number*, which took centuries to distill. That process is one important thread in the tapestry of mathematical history.

Very often, a spatial perception is purely geometrical or topological, involving similarity (having the same shape), connectivity (having holes or being solid), boundedness or infinitude, and the like. We can see the origins of these concepts in many aspects of everyday life that do not involve what one would call formal geometry. The perception of continuous magnitudes such as lengths, areas, volumes, weights, and time is different from the perception of multiple copies of a discrete object. The two kinds of perception work both independently and together to help a human being or animal cope with the physical world. Getting these two “draft horses” harnessed together as parts of a common subject called mathematics has led to a number of interesting problems to be solved.

2.1. Perception of shape by animals. Obviously, the ability to perceive shape is of value to an animal in determining what is or is not food, what is a predator, and so forth; and in fact the ability of animals to perceive space has been very well documented. One of the most fascinating examples is the ability of certain species of bees to communicate the direction and distance of sources of plant nectar by performing a dance inside the beehive. The pioneer in this work was Karl von

Frisch (1886–1982), and his work has been continued by James L. Gould and Carol Grant Gould (1995). The experiments of von Frisch left many interpretations open and were challenged by other specialists. The Goulds performed more delicately designed experiments which confirmed the bee language by deliberately misleading the bees about the food source. The bee will traverse a circle alternately clockwise and counterclockwise if the source is nearby. If it is farther away, the alternate traversals will spread out, resulting in a figure 8, and the dance will incorporate sounds and wagging. By moving food sources, the Goulds were able to determine the precision with which this communication takes place (about 25%). Still more intriguing is the fact that the direction of the food source is indicated by the direction of the axis of the figure 8, oriented relative to the sun if there is light and relative to the vertical if there is no light.

As another example, in his famous experiments on conditioned reflexes using dogs as subjects the Russian scientist Pavlov (1849–1936) taught dogs to distinguish ellipses of very small eccentricity from circles. He began by projecting a circle of light on the wall each time he fed the dog. Eventually the dog came to expect food (as shown by salivation) every time it saw the circle. When the dog was conditioned, Pavlov began to show the dog an ellipse in which one axis was twice as long as the other. The dog soon learned not to expect food when shown the ellipse. At this point the malicious scientist began making the ellipse less eccentric, and found, with fiendish precision, that when the axes were nearly equal (in a ratio of 8 : 9, to be exact) the poor dog had a nervous breakdown (Pavlov, 1928, p. 122).

2.2. Children's concepts of space. The most famous work on the development of mathematical concepts in children is due to Jean Piaget (1896–1980) of the University of Geneva, who wrote many books on the subject, some of which have been translated into English. Piaget divided the development of the child's ability to perceive space into three periods: a first period (up to about 4 months of age) consisting of pure reflexes and culminating in the development of primary habits, a second period (up to about one year) beginning with the manipulation of objects and culminating in purposeful manipulation, and a third period in which the child conducts experiments and becomes able to comprehend new situations. He categorized the primitive spatial properties of objects as proximity, separation, order, enclosure, and continuity. These elements are present in greater or less degree in any spatial perception. In the baby they come together at the age of about 2 months to provide recognition of faces. The human brain seems to have some special "wiring" for recognizing faces.

The interesting thing about these concepts is that mathematicians recognize them as belonging to the subject of topology, an advanced branch of geometry that developed in the late nineteenth and early twentieth centuries. It is an interesting paradox that the human ability to perceive shape depends on synthesizing various topological concepts; this progression reverses the pedagogical and historical ordering between geometry and topology. Piaget pointed out that children can make topological distinctions (often by running their hands over models) before they can make geometric distinctions. Discussing the perceptions of a group of 3-to-5-year-olds, Piaget and Inhelder (1967) stated that the children had no trouble distinguishing between open and closed figures, surfaces with and without holes, intertwined rings and separate rings, and so forth, whereas the seemingly simpler

relationships of geometry—distinguishing a square from an ellipse, for example were not mastered until later.

2.3. Geometry in arts and crafts. Weaving and knitting are two excellent examples of activities in which the spatial and numerical aspects of the world are combined. Even the sophisticated idea of a rectangular coordinate system is implicit in the placing of different-colored threads at intervals when weaving a carpet or blanket so that a pattern appears in the finished result. One might even go so far as to say that curvilinear coordinates occur in the case of sweaters.

Not only do arts and crafts *involve* the kind of abstract and algorithmic thinking needed in mathematics, their themes have often been inspired by mathematical topics. We shall give several examples of this inspiration in different parts of this book. At this point, we note just one example, which the author happened to see at a display of quilts in 2003. The quilt, shown on the cover of this book, embodies several interesting properties of the *Golden Ratio* $\Phi = (1 + \sqrt{5})/2$, which is the ratio of the diagonal of a pentagon to its side. This ratio is known to be involved in the way many trees and flowers grow, in the spiral shell of the chambered nautilus, and other places. The quilt, titled “A Number Called Phi,” was made by Mary Knapp of Watertown, New York. Observe how the quilter has incorporated the spiral connection in the sequence of nested circles and the rotation of each successive inscribed pentagon, as well as the phyllotaxic connection suggested by the vine.

Marcia Ascher (1991) has assembled many examples of rather sophisticated mathematics inspired by arts and crafts. The Bushoong people of Zaire make part of their living by supplying embroidered cloth, articles of clothing, and works of art to others in the economy of the Kuba chiefdom. As a consequence of this work, perhaps as preparation for it, Bushoong children amuse themselves by tracing figures on the ground. The rule of the game is that a figure must be traced without repeating any strokes and without lifting the finger from the sand. In graph theory this problem is known as the *unicursal tracing problem*. It was analyzed by the Swiss mathematician Leonhard Euler (1707–1783) in the eighteenth century in connection with the famous Königsberg bridge problem. According to Ascher, in 1905 some Bushoong children challenged the ethnologist Emil Torday (1875–1931) to trace a complicated figure without lifting his finger from the sand. Torday did not know how to do this, but he did collect several examples of such figures. The Bushoong children seem to learn intuitively what Euler proved mathematically: A unicursal tracing of a connected graph is possible if there are at most two vertices where an odd number of edges meet. The Bushoong children become very adept at finding such a tracing, even for figures as complicated as that shown in Fig. 1.

3. Symbols

We tend to think of symbolism as arising in algebra, since that is the subject in which we first become aware of it as a concept. The thing itself, however, is implanted in our minds much earlier, when we learn to talk. Human languages, in which sounds correspond to concepts and the temporal order or inflection of those sounds maps some relation between the concepts they signify, exemplify the process of abstraction and analogy, essential elements in mathematical reasoning. Language is, all by itself, ample proof that the symbolic ability of human beings is highly developed. That symbolic ability lies at the heart of mathematics.

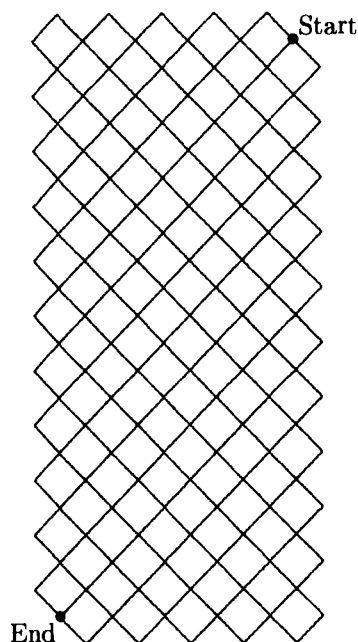


FIGURE 1. A graph for which a unicursal tracing is possible.

Once numbers have been represented symbolically, the next logical step would seem to be to introduce symbols for arithmetic operations or for combining the number symbols in other ways. However, this step may not be necessary for rapid computation, since mechanical devices such as counting rods, pebbles, counting boards, and the like can be used as analog computers. The operations performed using these methods can rise to a high level of sophistication without the need for any written computations. An example of the use of an automatic counting device is given by Ascher (1997) in a discussion of a system of divination used by the Malagasy of Madagascar, in which four piles of seeds are arranged in a column and the seeds removed from each pile two at a time until only one or two seeds remain. Each set of seeds in the resulting column can be interpreted as “odd” or “even.” After this procedure is performed four times, the four columns and four rows that result are combined in different pairs using the ordinary rules for adding odds and evens to generate eight more columns of four numbers. The accuracy of the generation is checked by certain mathematical consequences of the method used. If the results are satisfactory, the 16 sets of four odds and evens are used as an oracle for making decisions and ascribing causes to such events as illnesses.

The Malagasy system of divination bears a resemblance to the procedures described in the Chinese classic *I Ching* (*Permutation Classic*). In the latter, a set of 50 yarrow sticks is used, the first stick being laid down to begin the ceremony. One stick is then placed between the ring and small fingers of the left hand to represent the human race. The remaining 48 sticks are then divided without counting into two piles, and one pile held in each hand. Those in the right hand are then discarded four at a time until four or fewer remain. These are then transferred to the left hand, and the same reduction is applied to the other pile, so that at

the end, the left hand contains either five or nine sticks. After a sequence of such procedures, a final step begins with 32 or 36 or 40 sticks, and as a result the number of remaining sticks will be 24, 28, 32, or 36. This number is divided by four and the quotient determines the bottom row of the symbol to be used for divination. Six is called *lesser yang*, seven *greater ying*, eight *lesser ying*, and nine *greater yang*. The *ying* and *yang* are respectively female and male principles. The greater cases correspond to flux (tending to their opposites) and the lesser to stability. When this entire procedure has been carried out six times, the result is a stack of six symbols that can be interpreted according to the principles of divination. There are 64, that is, 2^6 , different possible stackings of *ying* and *yang*, all discussed in the *I Ching*, and the duality between stability and flux makes for 4096 possible symbols. One must beware of attaching too much significance to numerical coincidences, but it is intriguing that both Malagasy and Chinese forms of divination are based on the number four.¹

Divination seems to fulfill a nearly universal human desire to feel in control of the powerful forces that threaten human happiness and prosperity. It manifests itself in a variety of ways, as just shown by the examples of the Malagasy and the *I Ching*. We could also cite large parts of the Jewish *Kabbalah*, the mysticism of the Pythagoreans, and many others, down to the geometric logic of Ramon Lull (1232–1316), who was himself steeped in the *Kabbalah*. The variety of oracles that people have consulted for advice about the conduct of their lives—tarot cards, crystal balls, astrology, the entrails of animals and birds, palmistry, and the like—seems endless. For the purposes of this book, however, we shall be interested only in those aspects of divination that involve mathematics. Magic squares, for example, occur in both the *Kabbalah* and the *I Ching*. Although the author puts no stock whatsoever in the theories behind all this mysticism, it remains an important fact about human behavior over the centuries and deserves to be studied for that reason alone. But for now it is time to return to more prosaic matters.

Aids to computation, either tabular or mechanical, must be used to perform computations in some of the more cumbersome notational systems. Just imagine trying to multiply XLI by CCCIV! (However, Detlefsen and co-authors (1975) demonstrate that this task is not so difficult as it might seem.) Even to use the 28×19 table of dates of Easter discussed in Problem 6.26, the Slavic calculators had to introduce simplifications to accommodate the fact that dividing a four-digit number by a two-digit number was beyond the skill of many of the users of the table.

The earliest mathematical texts discuss arithmetical operations using everyday words that were probably emptied of their usual meaning. Students had to learn to generalize from a particular example to the abstract case, and many problems that refer to specific objects probably became archetypes for completely abstract reasoning, just as we use such expressions as “putting the cart before the horse” and “comparing apples and oranges” to refer to situations having no connection at all with horse-and-buggy travel or the harvesting of fruit. For example, problems of the

¹ Like all numbers, the number four is bound to occur in many contexts. One website devoted to spreading the lore found in the *I Ching* notes the coincidental fact that DNA code is written with four amino acids as its alphabet and rhapsodizes that “The sophistication of this method has not escaped modern interpretation, and the four-valued logic has been compared to the biochemistry of DNA amino acids. How a Neolithic shaman’s divination technique presaged the basic logic of the human genome is one of the ageless mysteries.”

type, “If 3 bananas cost 75 cents, how much do 7 bananas cost?” occur in the work of Brahmagupta from 1300 years ago. Brahmagupta named the three data numbers *argument* (3), *fruit* (75), and *requisition* (7). As another example, cuneiform tablets from Mesopotamia that are several thousand years old contain general problems that we would now solve using quadratic equations. These problems are stated as variants of the problem of finding the length and width of a rectangle whose area and perimeter are known. The mathematician and historian of mathematics B. L. van der Waerden (1903–1996) claimed that the words for *length* and *width* were being used in a completely abstract sense in these problems.

In algebra symbolism seems to have occurred for the first time in the work of the Greek mathematician Diophantus of Alexandria, who introduced the symbol ς for an unknown number. The Bakshali Manuscript, a document from India that may have been written within a century of the work of Diophantus, also introduces an abstract symbol for an unknown number. In modern algebra, beginning with the Muslim mathematicians more than a millennium ago, symbolism evolved gradually. Originally, the Arabic word for *thing* was used to represent the unknown in a problem. This word, and its Italian translation *cosa*, was eventually replaced by the familiar x most often used today. In this way an entire word was gradually pared down to a single letter that could be manipulated graphically.

4. Mathematical inference

Logic occurs throughout modern mathematics as one of its key elements. In the teaching of mathematics, however, the student generally learns all of arithmetic and the rules for manipulating algebraic expressions by rote. Any justification of these rules is purely experimental. Logic enters the curriculum, along with proof, in the study of Euclidean geometry. This sequence is not historical and may leave the impression that mathematics was an empirical science until the time of Euclid (ca. 300 BCE). Although one can imagine certain facts having been discovered by observation, such as the rule for comparing the area of a rectangle with the area of a square unit, there is good reason to believe that some facts were *deduced* from simpler considerations at a very early stage. The main reason for thinking so is that the conclusions reached by some ancient authors are not visually obvious.

4.1. Visual reasoning. As an example, it is immediately obvious that a diagonal divides a rectangle into two congruent triangles. If through any point on the diagonal we draw two lines parallel to the sides, these two lines will divide the rectangle into four rectangles. The diagonal divides two of these smaller rectangles into pairs of congruent triangles, just as it does the whole rectangle, thus yielding three pairs of congruent triangles, one large pair and two smaller pairs. It then follows (see Fig. 2) that the two remaining rectangles must have equal area, even though their shapes are different and *to the eye they do not appear to be equal*. For each of these rectangles is obtained by subtracting the two smaller triangles from the large triangle in which they are contained. When we find an ancient author mentioning that these two rectangles of different shape are equal, as if it were a well-known fact, we can be confident that this knowledge does not rest on an experimental or inductive foundation. Rather, it is the result of a combination of numerical and spatial reasoning.

Ancient authors often state *what* they know without saying *how* they know it. As the example just cited shows, we can be confident that the basis was not

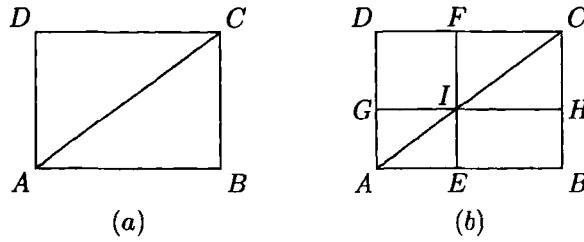


FIGURE 2. (a) The diagonal AC divides the rectangle $ABCD$ into congruent triangles ABC and CDA . (b) When the congruent pairs (AEI, IGA) and (IHC, CFI) are subtracted from the congruent pair (ABC, CDA) , the remainders (rectangles $EBHI$ and $GIFD$) must be equal.

always induction or experiment. Perminov (1997) points out that solutions of complicated geometric problems which can be shown to be correct are stated without proof—but apparently with absolute confidence—by the writers of the very earliest mathematical documents, such as the Rhind Papyrus from Egypt and cuneiform tablets from Mesopotamia. The fact that an author presents not merely a solution but a sequence of steps leading to that solution, and the fact that this solution can now be reconstructed and verified by mathematical reasoning, justify the conclusion that the result was arrived at through mathematical deduction, even though the author does not write out the details.

4.2. Chance and probability. Logic is concerned with getting conclusions that are as reliable as the premises. From a behavioral point of view, the human tendency to make inferences based on logic is probably hardwired and expressed as the same mechanism by which habits are formed. This same mechanism probably accounts for the metaphysical notion of *cause*. If A implies B , one feels that in some sense A *causes* B to be true. The dogs in Pavlov's experiments, described above, were given *total* reinforcement as they learned geometry and came to make associations based on the constant conjunction of a given shape and a given reward or lack of reward. In the real world, however, we frequently encounter a weaker type of cause, where A is usually, but not always, followed by B . For example, lightning is always followed by thunder, but if the lightning is very distant, the thunder will not be heard. The analog of this weaker kind of cause in conditioning is *partial reinforcement*. A classical example is a famous experiment of Skinner (1948), who put hungry pigeons in a cage and attached a food hopper to the cage with an automatic timer to permit access to the food at regular intervals. The pigeons at first engaged in aimless activity when not being fed, but tended to repeat whatever activity they happened to be doing when the food arrived, as if they made an association between the activity and the arrival of food. Naturally, the more they repeated a given activity, the more likely that activity was to be reinforced by the arrival of food. Since they were always hungry, it was not long before they were engaged full-time in an activity that they apparently considered an infallible food producer. This activity varied from one bird to another. One pigeon thrust its head into an upper corner of the cage; another made long sweeping movements with its

head; another tossed its head back; yet another made pecking motions toward the floor of the cage.

The difficulties that people, even mathematicians, have in understanding and applying probability can be seen in this example. For example, the human body has some capacity to heal itself. Like the automatic timer that eventually provided food to the pigeons, the human immune system often overcomes the disease. Yet sick people, like hungry pigeons, try various methods of alleviating their misery. The consequence is a wide variety of nostrums said to cure a cold or arthritis. One of the triumphs of modern mathematical statistics is the establishment of reliable systems of inference to replace the inferences that Skinner called "superstitious."

Modern logic has purged the concept of implication of all connection with the notion of cause. The statement "If Abraham Lincoln was the first President of the United States, then $2 + 2 = 4$ " is considered a true implication, even though Lincoln was not the first President and in any case his being such would have no *causal* connection with the truth of the statement " $2 + 2 = 4$." In standard logic the statement "If A is true, then B is true" is equivalent to the statement "Either B is true, or A is false, or both." Absolute truth or falsehood is not available in relation to the observed world, however. As a result, science must deal with propositions of the form "If A is true, then B is *highly probable*." One cannot infer from this statement that "If B is false, then A is *highly improbable*." For example, an American citizen, taken at random, is probably not a U.S. Senator. It does not follow that if a person *is* a U.S. Senator, that person is probably not an American citizen.

Questions and problems

1.1. At what point do you find it necessary to count in order to say how large a collection is? Can you look at a word such as *tendentious* and see immediately how many letters it has? The American writer Henry Thoreau (1817–1863) was said to have the ability to pick up exactly one dozen pencils out of a pile. Try as an experiment to determine the largest number of pencils you can pick up out of a pile without counting. The point of this exercise is to see where direct perception needs to be replaced by counting.

1.2. In what practical contexts of everyday life are the fundamental operations of arithmetic—addition, subtraction, multiplication, and division—needed? Give at least two examples of the use of each. How do these operations apply to the problems for which the theory of proportion was invented?

1.3. What significance might there be in the fact that there are three columns of notches on the Ishango Bone? What might be the significance of the numbers of notches in the three series?

1.4. Is it possible that the Ishango Bone was used for divination? Can you think of a way in which it could be used for this purpose?

1.5. Is it significant that one of the yarrow sticks is isolated at the beginning of each step in the Chinese divination procedure described above? What difference does this step make in the outcome?

1.6. Measuring a continuous object involves finding its ratio to some standard unit. For example, when you measure out one-third of a cup of flour in a recipe, you are

choosing a quantity of flour whose ratio to the standard cup is 1:3. Suppose that you have a standard cup without calibrations, a second cup of unknown size, and a large bowl. How could you determine the volume of the second cup?

1.7. Units of time, such as a day, a month, and a year, have ratios. In fact you probably know that a year is about $365\frac{1}{4}$ days long. Imagine that you had never been taught that fact. How would you—how did people originally—determine how many days there are in a year?

1.8. Why is a calendar needed by an organized society? Would a very small society (consisting of, say, a few dozen families) require a calendar if it engaged mostly in hunting, fishing, and gathering vegetable food? What if the principal economic activity involved following a reindeer herd? What if it involved tending a herd of domestic animals? Finally, what if it involved planting and tending crops?

1.9. Describe three different ways of measuring time, based on different physical principles. Are all three ways equally applicable to all lengths of time?

1.10. In what sense is it possible to know the *exact* value of a number such as $\sqrt{2}$? Obviously, if a number is to be known only by its decimal expansion, nobody does know and nobody ever will know the exact value of this number. What immediate practical consequences, if any, does this fact have? Is there any other sense in which one could be said to know this number *exactly*? If there are no direct consequences of being ignorant of its exact value, is there any practical value in having the *concept* of an exact square root of 2? Why not simply replace it by a suitable approximation such as 1.41421? Consider also other “irrational” numbers, such as π , e , and $\Phi = (1 + \sqrt{5})/2$. What is the value of having the *concept* of such numbers as opposed to approximate rational replacements for them?

1.11. Find a unicursal tracing of the graph shown in Fig. 1.

1.12. Does the development of personal knowledge of mathematics mirror the historical development of the subject? That is, do we learn mathematical concepts as individuals in the same order in which these concepts appeared historically?

1.13. Topology, which may be unfamiliar to you, studies (among other things) the mathematical properties of knots, which have been familiar to the human race at least as long as most of the subject matter of geometry. Why was such a familiar object not studied mathematically until the twentieth century?

1.14. One aspect of symbolism that has played a large role in human history is the mystical identification of things that exhibit analogous relations. The divination practiced by the Malagasy is one example, and there are hundreds of others: astrology, alchemy, numerology, tarot cards, palm reading, and the like, down to the many odd beliefs in the effects of different foods based on their color and shape. Even if we dismiss the validity of such divination (as the author does), is there any value for science in the development of these subjects?

1.15. What function does logic fulfill in mathematics? Is it needed to provide a psychological feeling of confidence in a mathematical rule or assertion? Consider, for example, any simple computer program that you may have written. What really gave you confidence that it worked? Was it your logical analysis of the operations involved, or was it empirical testing on an actual computer with a large variety of different input data?

1.16. Logic enters the mathematics curriculum in high-school geometry. The reason for introducing it at that stage is historical: Formal treatises with axioms, theorems, and proofs were a Greek innovation, and the Greeks were primarily geometers. There is no *logical* reason why logic is any more important in geometry than in algebra or arithmetic. Yet it seems that without the explicit statement of assumptions, the parallel postulate of Euclid (discussed in Chapter 10) would never have been questioned. Suppose things had happened that way. Does it follow that non-Euclidean geometry would never have been discovered? How important is non-Euclidean geometry, anyway? What other kinds of geometry do you know about? Is it necessary to be guided by axioms and postulates in order to discover or fully understand, say, the non-Euclidean geometry of a curved surface in Euclidean space? If it is not necessary, what is the value of an axiomatic development of such a geometry?

1.17. Perminov (1997, p. 183) presents the following example of tacit mathematical reasoning from an early cuneiform tablet. Given a right triangle ACB divided into a smaller triangle DEB and a trapezoid $ACED$ by the line DE parallel to the leg AC , such that EC has length 20, EB has length 30, and the trapezoid $ACED$ has area 320, what are the lengths AC and DE ? (See Fig. 3.) The author of the tablet very confidently computes these lengths by the following sequence of operations: (1) $320 \div 20 = 16$; (2) $30 \cdot 2 = 60$; (3) $60 + 20 = 80$; (4) $320 \div 80 = 4$; (5) $16 + 4 = 20 = AC$; (6) $16 - 4 = 12 = DE$. As Perminov points out, to present this computation with any confidence, you would have to know exactly what you are doing. What *was* this anonymous author doing?

To find out, fill in the reasoning in the following sketch. The author's first computation shows that a rectangle of height 20 and base 16 would have exactly the same area as the trapezoid. Hence if we draw the vertical line FH through the midpoint G of AD , and complete the resulting rectangles as in Fig. 3, rectangle $FCEI$ will have area 320. Since $AF = MI = FJ = DI$, it now suffices to find this common length, which we will call x ; for $AC = CF + FA = 16 + x$ and $DE = EI - DI = 16 - x$. By the principle demonstrated in Fig. 2, $JCED$ has the same area as $DKLM$, so that $DKLM + FJDI = DKLM + 20x$. Explain why $DKLM = 30 \cdot 2 \cdot x$, and hence why $320 = (30 \cdot 2 + 20) \cdot x$.

Could this procedure have been obtained experimentally?

1.18. A famous example of mathematical blunders committed by mathematicians (not statisticians, however) occurred some two decades ago. At the time, a very popular television show in the United States was called *Let's Make a Deal*. On that show, the contestant was often offered the chance to keep his or her current winnings, or to trade them for a chance at some other unknown prize. In the case in question the contestant had chosen one of three boxes, knowing that only one of them contained a prize of any value, but not knowing the contents of any of them. For ease of exposition, let us call the boxes A, B, and C, and assume that the contestant chose box A.

The emcee of the program was about to offer the contestant a chance to trade for another prize, but in order to make the program more interesting, he had box B opened, in order to show that it was empty. Keep in mind that the emcee *knew* where the prize was and would not have opened box B if the prize had been there. Just as the emcee was about to offer a new deal, the contestant asked to exchange the chosen box (A) for the unopened box (C) on stage. The problem posed to the

are equally likely. Hence the probability of this new event F is $1/2$. Thus, even though the mathematics of conditional probability is quite simple, it can be a subtle problem to describe just what event has occurred. Conclusion: To reason correctly in cases of conditional probability, *one must be very clear in describing the event that has occurred.*

1.21. Reinforcing the conclusion of Problem 1.20, exhibit the fallacy in the following “proof” that *lotteries are all dishonest.*

Proof. The probability of winning a lottery is less than one chance in 1,000,000 ($= 10^{-6}$). Since all lottery drawings are independent of one another, the probability of winning a lottery five times is less than $(10^{-6})^5 = 10^{-30}$. But this probability is far smaller than the probability of any conceivable event. Any scientist would disbelieve a report that such an event had actually been observed to happen. Since the lottery has been won five times in the past year, it must be that winning it is not a random event; that is, the lottery is fixed.

What is the event that has to occur here? Is it “Person A (specified in advance) wins the lottery,” or is it “At least one person in this population (of 30 million people) wins the lottery”? What is the difference between those two probabilities? (The same fallacy occurs in the probabilistic arguments purporting to prove that evolution cannot occur, based on the rarity of mutations.)

1.22. The relation between mathematical creativity and musical creativity, and the mathematical aspects of music itself are a fascinating and well-studied topic. Consider just the following problem, based on the standard tuning of a piano keyboard. According to that tuning, the frequency of the major fifth in each scale should be $3/2$ of the frequency of the base tone, while the frequency of the octave should be twice the base frequency. Since there are 12 half-tones in each octave, starting at the lowest A on the piano and ascending in steps of a major fifth, twelve steps will bring you to the highest A on the piano. If all these fifths are tuned properly, that highest A should have a frequency of $(\frac{3}{2})^{12}$ times the frequency of the lowest A. On the other hand, that highest A is seven octaves above the lowest, so that, if all the octaves are tuned properly, the frequency should be 2^7 times as high. The difference between these two frequency ratios, $7153/4096 \approx 1.746$ is called the *Pythagorean comma*. (The Greek word *komma* means a break or cutoff.) What is the significance of this discrepancy for music? Could you hear the difference between a piano tuned so that all these fifths are exactly right and a piano tuned so that all the octaves are exactly right? (The ratio of the discrepancy between the two ratios to either ratio is about 0.01%.)

1.23. What meaning can you make of the statement attributed to the French poet Sully (René François Armand) Prudhomme (1839–1907), “Music is the pleasure the soul experiences from counting without realizing it is counting”?

CHAPTER 2

Mathematical Cultures I

In Chapter 1 we looked at the origin of mathematics in the everyday lives of people. The evidence for the conclusions presented there is indirect, coming from archaeology, anthropology, and other studies not directly mathematical. Wherever there are written documents to refer to, we can know much more about what was done and why. The present chapter is a broad survey of the development of mathematics that arose spontaneously, as far as is known, in a number of cultures around the world. We are particularly interested in highlighting the motives for creating mathematics.

1. The motives for creating mathematics

As we saw in Chapter 1, a certain amount of numerical and geometric knowledge is embedded in the daily lives of people and even animals. Human beings at various times have developed more intricate and sophisticated methods of dealing with numbers and space, leading to arithmetic, geometry, and beyond. That kind of knowledge must be taught systematically if it is to be passed on from generation to generation and become a useful part of a civilization. Some group of people must devote at least a part of their time to learning and perhaps improving the knowledge that has been acquired. These people are professional mathematicians, although their primary activity may be commercial, administrative, or religious, and sometimes a combination of the three, as in ancient Egypt.

1.1. Pure versus applied mathematics. How does the mathematics profession arise? Nowadays people choose to enter this profession for a variety of reasons. Undoubtedly, an important motive is that they find mathematical ideas interesting to contemplate and work with; but if there were no way of making a living from having some expertise in the subject, the number of its practitioners would be far smaller. The question thus becomes: "Why are some people paid for solving mathematical problems and creating new mathematical knowledge?" Industry and government find uses for considerable numbers of mathematicians and statisticians. For those of a purer, less applied bent, the universities and offices of scientific research subsidized by governments provide the opportunity to do research on questions of pure mathematical interest without requiring an immediate practical application. This kind of research has been pursued for thousands of years, and it has always had some difficulty justifying itself. Here, for example, is a passage from Book 7 of Plato's *Republic* in which, discussing the education of the leaders of an ideal state, Socrates and Glaucon decide on four subjects that they must learn, namely arithmetic, geometry, astronomy, and music. These four subjects were later to form the famous *quadrivium* (fourfold path) of education in medieval Europe. After listing

arithmetic and geometry, they come to astronomy, and Socrates (the narrator of the dialogue) reports that Glaucon was in favor of including it:

"I certainly am," he said, "to have a clearer perception of time periods, both months and eras is proper in agriculture and navigation, and no less so in military strategy."

"You amuse me," I said, "in that you apparently fear the crowd, lest you seem to prescribe useless studies. But it is not an easy thing, it is difficult, to believe that some organ of the soul of each person is purified and refreshed in these studies, while it is lost and blinded by other pursuits; it is more to be preserved than ten thousand eyes, for by it alone is truth seen. Now some people will agree that what you are proposing is extremely good; but those who have never felt these things will regard you as having said nothing; for they see in them no profit worthy of the name."

Plato adopted the Pythagorean doctrine that there is a human faculty attuned to eternal truth and working through human reason. If there really is such a faculty, then of course mathematics is of high value for everyone. Plato, speaking here through Socrates, admits that some people seem to lack this faculty, so that it is difficult to believe in its universality. The difference in outlook between the two classes of people mentioned by Plato continues right down to the present time. Here, for example, is the view of the famous modern applied mathematician, R. W. Hamming (1915-1998), inventor of the Hamming codes, who if he believed at all in the "eye of the soul," at least did not believe it had a claim on public funds:

[T]he computing expert needs to be wary of believing much that he learns in his mathematics courses. . . [M]uch of modern mathematics is not related to science but. . . to the famous scholastic arguing of the Middle Ages. . . I believe it is important to make these distinctions. . . [T]he failure to do so has. . . caused government money appropriated for numerical analysis to be diverted to the art form of pure mathematics.

On the opposite side of the question is the following point of view, expressed by the famous British mathematician G. H. Hardy (1877-1947), who in 1940 wrote *A Mathematician's Apology*. After quoting an earlier address, in which he had said, "after all, the scale of the universe is large and, if we are wasting our time, the waste of the lives of a few university dons is no such overwhelming catastrophe,"¹ he gave what he considered the justification of his life:

I have never done anything "useful." No discovery of mine has made, or is likely to make, directly or indirectly, for good or ill, any difference to the amenity of the world. . . The case for my life, then, . . . is this: that I have added something to knowledge, and have helped others to add more; and that these somethings have a value which differs in degree only, not in kind, from that of the creations of the great mathematicians.

¹ From a cosmic point of view, no doubt, this is a cogent argument. From the point of view of that second group of people mentioned by Plato, however, it ignores the main question: Why should a person expect to receive a salary for doing work that others regard as useless?

As Plato said, one class of people will find this argument a sufficient justification for being a mathematician; another class will not. By reading the motivations given by other mathematicians for their work, the reader may either find arguments to convince that second group of people, or else come to agree with them and Hamming.

2. India

From archaeological excavations at Mohenjo Daro and Harappa on the Indus River in Pakistan it is known that an early civilization existed in this region for about a millennium starting in 2500 BCE. This civilization may have been an amalgam of several different cultures, since anthropologists recognize five different physical types among the human remains. Many of the artifacts that were produced by this culture have been found in Mesopotamia, evidence of trade between the two civilizations.

The Aryan civilization. The early civilization of these five groups of people disappeared around 1500 BCE, and its existence was not known in the modern world until 1925. The cause of its extinction is believed to be an invasion from the northwest by a sixth group of people, who spoke a language closely akin to early Greek. Because of their language these people are referred to as Aryans. The Aryans gradually expanded and formed a civilization of small kingdoms, which lasted about a millennium.

Sanskrit literature. The language of the Aryans became a literary language known as Sanskrit, in which great classics of literature and science have been written. Sanskrit thus played a role in southern Asia analogous to that of Greek in the Mediterranean world and Chinese in much of eastern Asia. That is, it provided a means of communication among scholars whose native languages were not mutually comprehensible and a basis for a common literature in which cultural values could be preserved and transmitted. During the millennium of Aryan dominance the spoken language of the people gradually diverged from written Sanskrit. Modern descendants of Sanskrit are Hindi, Gujarati, Bengali, and others. Sanskrit is the language of the *Mahabharata* and the *Ramayana*, two epic poems whose themes bear some resemblance to the Homeric epics, and of the *Upanishads*, which contain much of the moral teaching of Hinduism.

Among the most ancient works of literature in the world are the Hindu *Vedas*. The word means *knowledge* and is related to the English word *wit*. The composition of the *Vedas* began around 900 BCE, and additions continued to be made to them for several centuries. Some of these *Vedas* contain information about mathematics, conveyed incidentally in the course of telling important myths.

Hindu religious reformers. Near the end of the Aryan civilization, in the second half of the sixth century BCE, two figures of historical importance arise. The first of these was Gautama Buddha, the heir to a kingdom near the Himalayas, whose spiritual journey through life led to the principles of Buddhism. The second leader, Mahavira, is less well known but has some importance for the history of mathematics. Like his contemporary Buddha, he began a reform movement within Hinduism. This movement, known as Jainism, still has several million adherents in India. It is based on a metaphysic that takes very seriously what is known in some Western ethical systems as the *chain of being*. Living creatures are ranked

according to their awareness. Those having five senses are the highest, and those having only one sense are the lowest.

Islam in India. The amazingly rapid Muslim expansion from the Arabian desert in the seventh century brought Muslim invaders to India by the early eighth century. The southern valley of the Indus River became a province of the huge Umayyad Empire, but the rest of India preserved its independence, as it did 300 years later when another Muslim people, the Turks and Afghans, invaded. Still, the contact was enough to bring certain Hindu works, including the Hindu numerals, to the great center of Muslim culture in Baghdad. The complete and destructive conquest of India by the Muslims under Timur the Lame came at the end of the fourteenth century. Timur did not remain in India but sought new conquests; eventually he was defeated by the Ming dynasty in China. India was desolated by his attack and was conquered a century later by Akbar the Lion, a descendant of both Genghis Khan and Timur the Lame and the first of the Mogul emperors. The Mogul Empire lasted nearly three centuries and was a time of prosperity and cultural resurgence. One positive effect of this second Muslim expansion was a further exchange of knowledge between the Hindu and Muslim worlds. Interestingly, the official administrative language used for Muslim India was neither Arabic nor an Indian language; it was Persian.

British rule. During the seventeenth and eighteenth centuries British and French trading companies were in competition for the lucrative trade with the Mogul Empire. British victories during the Seven Years War left Britain in complete control of this trade. Coming at the time of Mogul decline due to internal strife among the Muslims and continued resistance on the part of the Hindus, this trade opened the door for the British to make India part of their empire. British colonial rule lasted nearly 200 years, coming to an end only after World War II. British rule made it possible for European scholars to become acquainted with Hindu classics of literature and science. Many Sanskrit works were translated into English in the early nineteenth century and became part of the world's science and literature.

We can distinguish three periods in the development of mathematics in the Indian subcontinent. The first period begins around 900 BCE with individual mathematical results forming part of the Vedas. The second begins with systematic treatises concerned mostly with astronomy but containing explanations of mathematical results, which appear in the second century CE. These treatises led to continuous progress for 1500 years, during which time much of algebra, trigonometry, and certain infinite series that now form part of calculus were discovered, a century or more before Europeans developed calculus. In the third stage, which began during the two centuries of British rule, this Hindu mathematics came to be known in the West, and Indian mathematicians began to work and write in the modern style of mathematics that is now universal.

2.1. The *Sulva Sūtras*. In the period from 800 to 500 BCE a set of verses of geometric and arithmetic content were written and became part of the *Vedas*.²

² Most of this discussion of the *Sulva Sūtras* is based on the work of Srinivasiengar (1967), which gives a clear exposition but contains statements that are rather alarming for one who is forced to rely on a secondary source. For example, on p. 6 we learn that the unit of length known as the *vyayam* was "about 96 inches," and "possibly this represented the height of the average man in those days." Indeed. Where, Mr. Srinivasiengar, have archaeologists discovered 8-foot-tall human skeletons?

These verses are known collectively as the *Sulva Sūtras* or *Sulba Sūtras*. The name means *Cord Rules* and probably reflects the use of a stretched rope or cord as a way of measuring length. The root *sulv* originally meant *to measure* or *to rule*, although it also has the meaning of a cord or rope; *sūtra* means *thread* or *cord*, a common measuring instrument. In the case of the *Vedas* the objects being measured with the cords were altars. The maintenance of altar fires was a duty for pious Hindus, and because Hinduism is polytheistic, it was necessary to consider how elaborate and large the fire dedicated to each deity was to be. This religious problem led to some interesting problems in arithmetic and geometry.

Two scholars who studied primarily the Sanskrit language and literature made important contributions to mathematics. Pingala, who lived around 200 BCE, wrote a treatise known as the *Chandaśśūtra*, containing one very important mathematical result, which, however, was stated so cryptically that one must rely on a commentary written 1200 years later to know what it meant. Later, a fifth-century scholar named Panini standardized the Sanskrit language, burdening it with some 4000 grammatical rules that make it many times more difficult to learn than any other Indo-European language. In the course of doing so, he made extensive use of combinatorics and the kind of abstract reasoning that we associate with algebra. These subjects set the most ancient Hindu mathematics apart from that of other nations.

2.2. Buddhist and Jaina mathematics. As with any religion that encourages quiet contemplation and the renunciation of sensual pleasure, Jainism often leads its followers to study mathematics, which provides a different kind of pleasure, one appealing to the mind. There have always been some mathematicians among the followers of Jainism, right down to modern times, including one in the ninth century bearing the same name as the founder of Jainism. The early work of Jaina mathematicians is notable for algebra (the *Sthānanga Sūtra*, from the second century BCE), for its concentration on topics that are essentially unique to early Hindu mathematics, such as combinatorics (the *Bhagabati Sūtra*, from around 300 BCE), and for speculation on infinite numbers (the *Anuyoga Dvāra Sūtra*, probably from the first century BCE). Buddhist monks were also very fond of large numbers, and their influence was felt when Buddhism spread to China in the sixth century CE.

2.3. The Bakshali Manuscript. A birchbark manuscript unearthed in 1881 in the village of Bakshali, near Peshawar, is believed by some scholars to date from the seventh century CE, although Sarkor (1982) believes it cannot be later than the end of the third century, since it refers to coins named *dīnāra* and *dramma*, which are undoubtedly references to the Greek coins known as the denarius and the drachma, introduced into India by Alexander the Great. These coins had disappeared from use in India by the end of the third century. The Bakshali Manuscript contains some interesting algebra, which is discussed in Chapter 14.

2.4. The *siddhantas*. During the second, third, and fourth centuries CE, Hindu scientists compiled treatises on astronomy known as *siddhantas*. The word *siddhanta* means a system.³ One of these treatises, the *Sūrya Siddhanta* (System of

³ A colleague of the author suggested that this word may be cognate with the Greek *idōn*, the aorist participle of the verb meaning *see*.

the Sun), from the late fourth century, has survived intact. Another from approximately the same time, the *Paulisha Siddhanta*, was frequently referred to by the Muslim scholar al-Biruni (973–1048). The name of this treatise seems to have been bestowed by al-Biruni, who says that the treatise was written by an Alexandrian astrologer named Paul.

2.5. Aryabhata I. With the writing of treatises on mathematics and astronomy, we at last come to some records of the motives that led people to create Hindu mathematics, or at least to write expositions of it. A mathematician named Aryabhata, the first of two mathematicians bearing that name, lived in the late fifth and early sixth centuries at Kusumapura (now Pataliputra, a village near the city of Patna) and wrote a book called the *Aryabhatiya*. This work had been lost for centuries when it was recovered by the Indian scholar Bhau Daji (1822–1874) in 1864. Scholars had known of its existence through the writings of commentators and had been looking for it. Writing in 1817, the English scholar Henry Thomas Colebrooke (1765–1837), who translated other Sanskrit mathematical works into English, reported, “A long and diligent research of various parts of India has, however, failed of recovering any part of the. . . *Algebra* and other works of Aryabhata.” Ten years after its discovery the *Aryabhatiya* was published at Leyden and attracted the interest of European and American scholars. It consists of 123 stanzas of verse, divided into four sections, of which the first, third, and fourth are concerned with astronomy and the measurement of time.

Like all mathematicians, Aryabhata I was motivated by intellectual interest. This interest, however, was closely connected with his Hindu piety. He begins the *Aryabhatiya* with the following tribute to the Hindu deity:

Having paid reverence to Brahman, who is one but many, the true deity, the Supreme Spirit, Aryabhata sets forth three things: mathematics, the reckoning of time, and the sphere. [Clark, 1930, p. 1]

The translator adds phrases to explain that Brahman is one as the sole creator of the universe, but is many via a multitude of manifestations.

Aryabhata then continues his introduction with a list of the astronomical observations that he will be accounting for and concludes with a promise of the reward awaiting the one who learns what he has to teach:

Whoever knows this *Dasagitika Sutra* which describes the movements of the Earth and the planets in the sphere of the asterisms passes through the paths of the planets and asterisms and goes to the higher Brahman. [Clark, 1930, p. 20]

As one can see, students in Aryabhata's culture had an extra reason to study mathematics and astronomy, beyond the concerns of practical life and the pleasures of intellectual edification. Learning mathematics and astronomy helped to advance the soul through the cycle of births and rebirths that Hindus believed in.

After setting out his teaching on the three subjects, Aryabhata concludes with a final word of praise for the Hindu deity and invokes divine endorsement of his labors:

By the grace of God the precious sunken jewel of true knowledge has been rescued by me, by means of the boat of my own knowledge, from the ocean which consists of true and false knowledge. He who disparages this universally true science of astronomy, which formerly was revealed by Svayambhu and is now described by me in this Aryabhata, loses his good deeds and his long life. [Clark, 1930, p. 81]

2.6. Brahmagupta. The establishment of research centers for astronomy and mathematics at Kusumapura and Ujjain produced a succession of good mathematicians and mathematical works for many centuries after Aryabhata I. About a century after Aryabhata I another Hindu mathematician, Brahmagupta, was born in the city of Sind, now in Pakistan. He was primarily an astronomer, but his astronomical treatise, the *Brahmasphutasiddhanta* (literally *The Corrected Brahma Siddhanta*), contains several chapters on computation (*Ganita*). The Hindu interest in astronomy and mathematics continued unbroken for several centuries, producing important work on trigonometry in the tenth century.

2.7. Bhaskara II. Approximately 500 years after Brahmagupta, in the twelfth century, the mathematician Bhaskara, the second of that name, was born on the site of the modern city of Bijapur. He is the author of the *Siddhanta Siromani*, in four parts, a treatise on algebra and geometric astronomy. Only the first of these parts, known as the *Lilavati*, and the second, known as the *Vija Ganita*,⁴ concern us here. Bhaskara says that his work is a compendium of knowledge, a sort of textbook of astronomy and mathematics. The name *Lilavati*, which was common among Hindu women, seems to have been a fancy of Bhaskara himself. Many of the problems are written in the form of puzzles addressed to this Lilavati.

Bhaskara II apparently wrote the *Lilavati* as a textbook to form part of what we would call a liberal education. His introduction reads as follows:

Having bowed to the deity, whose head is like an elephant's, whose feet are adored by gods; who, when called to mind, relieves his votaries from embarrassment; and bestows happiness on his worshippers; I propound this easy process of computation, delightful by its elegance, perspicuous with words concise, soft and correct, and pleasing to the learned. [Colebrooke, 1817, p. 1]

As a final advertisement at the end of his book, Bhaskara extols the pleasure to be derived from learning its contents:

Joy and happiness is indeed ever increasing in this world for those who have *Lilavati* clasped to their throats, decorated as the members are with neat reduction of fractions, multiplication, and involution, pure and perfect as are the solutions, and tasteful as is the speech which is exemplified. [Colebrooke, 1817, p. 127]

⁴ This Sanskrit word means literally *source computation*. It is compounded from the Sanskrit root *vij-* or *bij-*, which means *seed*. As discussed in Chapter 13, the basic idea of algebra is to find one or more numbers (the "source") knowing the result of operating on them in various ways. The word is usually translated as *algebra*.

The *Vija Ganita* consists of nine chapters, in the last of which Bhaskara tells something about himself and his motivation for writing the book:

On earth was one named Maheswara, who followed the eminent path of a holy teacher among the learned. His son Bhaskara, having from him derived the bud of knowledge, has composed this brief treatise of elemental computation. As the treatises of algebra [*vija ganita*] by Brahmagupta, Shidhara and Padmanabha are too diffusive, he has compressed the substance of them in a well-reasoned compendium for the gratification of learners... to augment wisdom and strengthen confidence. Read, do read, mathematician, this abridgement, elegant in style, easily understood by youth, comprising the whole essence of computation, and containing the demonstration of its principles, replete with excellence and void of defect. [Colebrooke, 1817, pp. 275–276]

2.8. Muslim India. Indian mathematical culture reflects the religious division between the Muslim and Hindu communities to some extent. The Muslim conquest brought Arabic and Persian books on mathematics to India. Some of these works were translated from ancient Greek, and among them was Euclid's *Elements*. These translations of later editions of Euclid contained certain obscurities and became the subject of commentaries by Indian scholars. Akbar the Lion decreed a school curriculum for Muslims that included three-fourths of what was known in the West as the quadrivium. Akbar's curriculum included arithmetic, geometry, and astronomy, leaving out only music.⁵ Details of this Indian Euclidean tradition are given in the paper by de Young (1995).

2.9. Indian mathematics in the colonial period and after. One of the first effects of British rule in India was to acquaint European scholars with the treasures of Hindu mathematics described above. It took a century before the British colonial rulers began to establish universities along European lines in India. According to Varadarajan (1983), these universities were aimed at producing government officials, not scholars. As a result, one of the greatest mathematical geniuses of all time. Srinivasa Ramanujan (1887–1920), was not appreciated and had to appeal to mathematicians in Britain to gain a position that would allow him to develop his talent. The necessary conditions for producing great mathematics were present in abundance, however, and the establishment of the Tata Institute in Bombay (now Mumbai) and the Indian Statistical Institute in Calcutta were important steps in this direction. After Indian independence was achieved, the first prime minister, Jawaharlal Nehru (1889–1964), made it a goal to achieve prominence in science. This effort has been successful in many areas, including mathematics. The names of Komaravolu Chandrasekharan (b. 1920), Harish-Chandra (1923–1983), and others have become celebrated the world over for their contributions to widely diverse areas of mathematics.

⁵ The quadrivium is said to have been proposed by Archytas (ca. 428–350 BCE), who lived in southern Italy and apparently communicated it to Plato when the latter was there to consult with the ruler of Syracuse; Plato incorporated it in his writings on education, as discussed in Sect. 1.

Srinivasa Ramanujan. The topic of power series is one in which Indian mathematicians had anticipated some of the discoveries in seventeenth- and eighteenth-century Europe. It was a facility with this technique that distinguished Ramanujan, who taught himself mathematics after having been refused admission to universities in India. After publishing a few papers, starting in 1911, he was able to obtain a stipend to study at the University of Madras. In 1913 he took the bold step of communicating some of his results to G.H. Hardy. Hardy was so impressed by Ramanujan's ability that he arranged for Ramanujan to come to England. Thus began a collaboration that resulted in seven joint papers with Hardy, while Ramanujan alone was the author of some 30 others. He rediscovered many important formulas and made many conjectures about functions such as the hypergeometric function that are represented by power series.

Unfortunately, Ramanujan was in frail health, and the English climate did not agree with him. Nor was it easy for him to maintain his devout Hindu practices so far from his normal Indian diet. He returned to India in 1919, but succumbed to illness the following year. Ramanujan's notebooks have been a subject of continuing interest to mathematicians. Hardy passed them on to G.N. Watson (1886–1965), who published a number of "theorems stated by Ramanujan." The full set of notebooks was published in the mid-1980s (see Berndt, 1985).

3. China

The name *China* refers to a region unified under a central government but whose exact geographic extent has varied considerably over the 4000 years of its history. To frame our discussion we shall sometimes refer to the following dynasties:⁶

The Shang Dynasty (sixteenth to eleventh centuries BCE). The Shang rulers controlled the northern part of what is now China and had an extensive commercial empire.

The Zhou Dynasty (eleventh to eighth centuries BCE). The Shang Dynasty was conquered by people from the northwest known as the Zhou. The great Chinese philosophers known in the West as Confucius, Mencius, and Lao-Tzu lived and taught during the period of disorder that came after the decay of this dynasty.

The Period of Warring States (403–221 BCE) and the *Qin Dynasty* (221–206 BCE). Warfare was nearly continuous in the fourth and third centuries BCE, but in the second half of the third century the northwestern border state of Qin gradually defeated all of its rivals and became the supreme power under the first Qin emperor. The name *China* is derived from the Qin.

The Han Dynasty (206 BCE–220 CE). The empire was conquered shortly after the death of the great emperor by people known as the Han, who expanded their control far to the south, into present-day Viet Nam, and established a colonial rule in the Korean peninsula. Contact with India during this dynasty brought Buddhism to China for the first time. According to Mikami (1913, pp. 57–58), mathematical and astronomical works from India were brought to China and studied. Certain topics, such as combinatorics, are common to both Indian and Chinese treatises, but "there

⁶ Because of total ignorance of the Chinese language, the author is forced to rely on translations of all documents. We shall write Chinese words in the Latin alphabet but not strive for consistency among the various sources that use different systems. We shall also omit the accent marks used to indicate the pitch of the vowels, since these cannot be pronounced by foreigners without special training.

is nothing positive that serves as an evidence of any actual Indian influence upon the Chinese mathematics."

The Tang Dynasty (seventh and eighth centuries). The Tang Dynasty was a period of high scholarship, in which, for example, block printing was invented.

The Song Dynasty (960–1279). The period of disorder after the fall of the Tang Dynasty ended with the accession of the first Song emperor. Confucianism underwent a resurgence in this period, supplementing its moral teaching with metaphysical speculation. As a result, a large number of scientific treatises on chemistry, zoology, and botany were written, and the Chinese made great advances in algebra.

The Mongol conquest and the closing of China. The Song Dynasty was ended in the thirteenth century by the Mongol conquest under the descendants of Genghis Khan, whose grandson Kublai Khan was the first emperor of the dynasty known to the Chinese as the Yuan. As the Mongols were Muslims, this conquest brought China into contact with the intellectual achievements of the Muslim world. Knowledge flowed both ways, of course, and the sophisticated Chinese methods of root extraction seem to be reflected in the works of later Muslim scholars, such as the fifteenth-century mathematician al-Kashi. The vast Mongol Empire facilitated East–West contacts, and it was during this period that Marco Polo (1254–1324) made his famous voyage to the Orient.

The Ming Dynasty (fourteenth to seventeenth centuries). While the Mongol conquest of Russia lasted 240 years, the Mongols were driven out of China in less than a century by the first Ming emperor. During the Ming Dynasty, Chinese trade and scholarship recovered rapidly. The effect of the conquest, however, was to encourage Chinese isolationism, which became the official policy of the later Ming emperors during the period of European expansion. The first significant European contact came in the year 1582, when the Jesuit priest Matteo Ricci (1552–1610) arrived in China. The Jesuits were particularly interested in bringing Western science to China to aid in converting the Chinese to Christianity. They persisted in these efforts despite the opposition of the emperor. The Ming Dynasty ended in the mid-seventeenth century with conquest by the Manchus.

The Ching (Manchu) Dynasty (1644–1911). After two centuries of relative prosperity the Ching Dynasty suffered from the depredations of foreign powers eager to control its trade. Perhaps the worst example was the Opium War of 1839–1842, fought by the British in order to gain control of the opium trade. From that time on Manchu rule declined. In 1900 the Boxer Rebellion against the Western occupation was crushed and the Chinese were forced to pay heavy reparations. In 1911 the government disintegrated entirely, and a republic was declared.

The twentieth century. The establishment of a republic in China did not quell the social unrest, and there were serious uprisings for several decades. China suffered badly from World War II, which began with a Japanese invasion in the 1930s. Although China was declared one of the major powers when the United Nations was formed in 1946, the Communist revolution of 1949 drove the ruler Chiang Kai-Shek to the island of Taiwan. The United States recognized only the government in Taiwan until the early 1970s. Then, bowing to the inevitable, it recognized the Communist government and did not use its veto power when that government replaced the Taiwanese government on the Security Council of the United Nations. At present, the United States recognizes two Chinese governments, one in Beijing and one on Taiwan. This view is contradicted by the government in Beijing, which

claims the authority to rule Taiwan. The American mathematician Walter Feit (1930–2004) visited China in May 1976 and reported that there was a heavy emphasis on combining mathematical theory with practice to solve social problems (Feit, 1977). At first zealous in adhering to the Maoist version of Marxism, the Chinese Communist Party undertook major reforms during the 1990s and has been moving in the direction of a more market-driven economy since then, although democracy seems to be slow in arriving. China is now engaged in extensive cultural and commercial exchanges with countries all over the world and hosted the International Congress of Mathematicians in 2002. Its mathematicians have made outstanding contributions to the advancement of mathematics, and Chinese students are eagerly welcomed at universities in nearly every country.

3.1. Works and authors. Mathematics became a recognized and respected area of intellectual endeavor in China more than 2000 years ago. That its origins are at least that old is established by the existence of books on mathematics, at least one of which was probably written before the order of the Emperor Shih Huang Ti in 213 BCE that all books be burned.⁷ A few books survived or were reconstituted after the brief reign of Shih Huang-Ti, among them the mathematical classic just alluded to. This work and three later ones now exist in English translation, with commentaries to provide the proper context for readers who are unfamiliar with the history and language of China. Under the Tang dynasty a standardized educational system came into place for the training of civil servants, based on literary and scientific classics, and the works listed below became part of a mathematical curriculum known as the *Suan Jing Shishu* (*Ten Canonical Mathematical Classics*—there are actually 12 of them). Throughout this long period mathematics was cultivated together with astronomy both as an art form and for their practical application in the problem of obtaining an accurate lunisolar calendar. In addition, many problems of commercial arithmetic appear in the classic works.

The Zhou Bi Suan Jing. The early treatise mentioned above, the *Zhou Bi Suan Jing*, has been known in English as the *Arithmetic Classic of the Gnomon and the Circular Paths of Heaven*. A recent, very thorough study and English translation has been carried out by Christopher Cullen of the University of London (1996). According to Cullen, the title *Zhou Bi* could be rendered as *The Gnomon of the Zhou*. The phrase *suan jing* occurs in the titles of several early mathematical works; it means *mathematical treatise* or *mathematical manual*. According to a tradition, the *Zhou Bi Suan Jing* was written during the Western Zhou dynasty, which overthrew the earlier Shang dynasty around 1025 BCE and lasted until 771 BCE. Experts now believe, however, that the present text was put together during the Western Han dynasty, during the first century BCE, and that the commentator Zhao Shuang, who wrote the version we now have, lived during the third century CE, after the fall of the Han dynasty. However, the astronomical information in the book could only have been obtained over many centuries of observation, and therefore must be much earlier than the writing of the treatise.

As the traditional title shows, the work is concerned with astronomy and surveying. The study of astronomy was probably regarded as socially useful in two ways: (1) It helped to regulate the calendar, a matter of great importance when rituals were to be performed; (2) it provided a method of divination (astrology),

⁷ The Emperor was not hostile to learning, since he did not forbid the *writing* of books. Apparently, he just wanted to be remembered as the emperor in whose reign everything began.

also of importance both for the individual and for the state. Surveying is of use in any society where it is necessary to erect large structures such as dams and bridges, and where land is often flooded, requiring people to abandon their land holdings and reclaim them later. These considerations at least provide a reason for people to regard mathematics as useful in practice. However, the preface, written by the commentator Zhao Shuang, gives a different version of the motive for compiling this knowledge. Apparently a student of traditional Chinese philosophy, he had realized that it was impossible to understand fully all the mysteries of the changing universe. He reports that he had looked into this work while convalescing from an illness and had been so impressed by the acuity of the knowledge it contained that he decided to popularize it by writing commentaries to help the reader over the hard parts, saying, "Perhaps in time gentlemen with a taste for wide learning may turn their attention to this work" (Cullen, 1996, p. 171).

Here we see mathematics being praised simply because it confers understanding where ignorance would otherwise be; it is regarded as a liberal art, to be studied by a leisured class of *gentlemen* scholars, people fortunate enough to be free of the daily grind of physical labor that was the lot of the majority of people in all countries until very recent times.

The Jiu Zhang Suanshu. Another ancient Chinese treatise, the *Jiu Zhang Suanshu*, meaning *Nine Chapters on the Mathematical Art*,⁸ has been partly translated into English, with commentary, by Lam (1994). A corrected and commented edition was published in Chinese in 1992, assembled by Guo (1992). This work has claim to be called *the* classic Chinese mathematical treatise. It reflects the level of mathematics in China in the later Han dynasty, around the year 100 CE. The nine chapters that give this monograph its name contain 246 applied problems of a sort useful in teaching how to handle arithmetic and elementary algebra and how to apply them in commercial and administrative work. Unfortunately, these chapters have no prefaces in which the author explains their purpose, and so we must assume that the purpose was the obvious one of training people engaged in surveying, administration, and trade. Some of the problems have an immediately practical nature, explaining how to find areas, convert units of length and area, and deal with fractions and proportions. Yet when we analyze the algebraic parts of this work, we shall see that it contains impractical puzzle-type problems leading to systems of linear equations and resembling problems that have filled up algebra books for centuries. Such problems are apparently intended to train the mind in algebraic thinking.

The Sun Zi Suan Jing. The most elementary of the early treatises is the *Sun Zi Suan Jing*, or *Mathematical Classic of Sun Zi*, even though its date is several centuries later than the *Jiu Zhang Suanshu*. This work begins with a preface praising the universality of mathematics for its role in governing the lives of all creatures, and placing it in the context of Chinese philosophy and among the six fundamental arts (propriety, music, archery, charioteership, calligraphy, and mathematics).

The preface makes it clear that mathematics is appreciated as both a practical skill in life and as an intellectual endeavor. The practicality comes in the use of compasses and gnomons for surveying and in the use of arithmetic for computing weights and measures. The intellectual skill, however, is emphasized. Mathematics

⁸ Chinese titles are apparently very difficult to render in English. Martzloff (1994) translates this title as *Computational Prescriptions in Nine Chapters*.

is valued because it trains the mind. "If one neglects its study, one will not be able to achieve excellence and thoroughness" (Lam and Ang, 1992, p. 151).

As in the quotation from the commentary on the *Zhou Bi Suan Jing*, we find that an aura of mystery and "elitism" surrounds mathematics. It is to be pursued by a dedicated group of initiates, who expect to be respected for learning its mysteries. At the same time, it has a practical value that is also respected.

Liu Hui. The *Hai Dao Suan Jing*. The fall of the Han Dynasty in the early third century gave rise to three separate kingdoms in the area now known as China. The north-central kingdom is known as the Kingdom of Wei. There, in the late third century CE, a mathematician named Liu Hui (ca. 220–280) wrote a commentary on the final chapter of the *Jiu Zhang Suanshu*. This chapter is devoted to the theorem we know as the Pythagorean theorem, and Liu Hui's book, the *Hai Dao Suan Jing* (*Sea Island Mathematical Classic*), shows how to use pairs of similar right triangles to measure inaccessible distances. The name of the work comes from the first problem in it, which is to find the height of a mountain on an offshore island and the distance to the base of the mountain. The work consists of nine problems in surveying that can be solved by the algebraic techniques practiced in China at the time. A translation of these problems, a history of the text itself, and commentary on the mathematical techniques can be found in the paper by Ang and Swetz (1986).

Liu Hui wrote a preface explaining that because of the burning of the books 400 years earlier, the few ancient texts still around had deteriorated, but that a minister of agriculture named Zhang Cang had produced a revised and corrected edition. However, most historians think that the *Jiu Zhang Suanshu* was written around 200 BC, after Shih Huang Ti ordered the burning of the books.

Zu Chongzhi and Zu Geng. According to Li and Du (1987, pp. 80–82), fifth-century China produced two outstanding mathematicians, father and son. Zu Chongzhi (429–501) and his son Zu Geng (ca. 450–520) were geometers who devised a method resembling what is now called *Cavalieri's principle* for calculating volumes bounded by curved surfaces. The elder Zu was also a numerical analyst, who wrote a book on approximation entitled *Zhui Shu* (*Method of Interpolation*), which became for a while part of the classical curriculum. However, this book was apparently regarded as too difficult for nonspecialists, and it was dropped from the curriculum and lost. Zu Geng continued working in the same area as his father and had a son who also became a mathematician.

Yang Hui. We now leave a considerable (700-year) gap in the story of Chinese mathematics. The next mathematician we wish to mention is Yang Hui (ca. 1238–1298), the author of a number of mathematical texts. According to Li and Du (1987, pp. 110, 115), one of these was *Xiangjie Jiuzhang Suan Fa* (*Detailed Analysis of the Mathematical Rules in the Jiu Zhang Suanshu*), a work of 12 chapters, one on each of the nine chapters of the *Jiu Zhang Suanshu*, plus three more containing other methods and more advanced analysis. In 1274 and 1275 he wrote two other works, which were later collected in a single work called the *Yang Hui Suan Fa* (*Yang Hui's Computational Methods*). In these works he discussed not only mathematics, but also its pedagogy, advocating real understanding over rote learning.

Zhu Shijie. Slightly later than Yang Hui, but still contemporary with him, was Zhu Shijie (ca. 1260–1320). He was still a young man in 1279, when China was united

by the Mongol emperor Kublai Khan and its capital established at what is now Beijing. Unification of the country enabled Zhu Shijie to travel more widely than had previously been possible. As a result, his *Suan Shu Chimeng* (*Introduction to Mathematical Studies*), although based on the *Jiu Zhang Suanshu*, went beyond it, discussing the latest methods of Chinese algebra. The original of this book was lost in Chinese, but a Korean version was later exported to Japan, where it had considerable influence. Eventually, a translation back from Korean into Chinese was made in the nineteenth century. According to Zharov (2001), who analyzed four fragments from this work, it shows some influence of Hindu or Arabic mathematics in its classification of large numbers. Zharov also proposed that the title be translated as “Explanation of some obscurities in mathematics” but says that his Chinese colleagues argued that the symbols for *chi* and *meng* were written as one and should be considered a single concept.

The Suan Fa Tong Zong of Cheng Dawei. A later work, the *Suan Fa Tong Zong* (*Treatise on Arithmetic*) by Cheng Dawei (1533–1606), was published in 1592. This book is well described by its title. It contains a systematic treatment of the kinds of problems handled in traditional Chinese mathematics, and at the end has a bibliography of some 50 other works on mathematics. The author, according to one of his descendents, was fascinated by books discussing problems on fields and grain, and assembled this book of problems over a lifetime of purchasing such books. Like the book of Zhu Shijie, Cheng Dawei’s book had a great influence on the development of mathematics in Korea and Japan. According to Li and Du (1987, p. 186), Cheng Dawei left a record of his mathematical studies, saying that he had been involved in travel and trade when young and had sought teachers everywhere he went. He retired from this profession while still young and spent 20 years consolidating and organizing his knowledge, so that “finally I rooted out the false and the nonsensical, put all in order, and made the text lucid.”

3.2. China’s encounter with Western mathematics. Jesuit missionaries who entered China during the late sixteenth century brought with them some mathematical works, in particular Euclid’s *Elements*, the first six books of which the missionary Matteo Ricci and the Chinese scholar Xu Guangchi (1562–1633) translated into Chinese (Li and Du, 1987, p. 193). The version of Euclid that they used, a Latin translation by the German Jesuit Christopher Clavius (1538–1612) bearing the title *Euclidis elementorum libri XV* (*The Fifteen Books of Euclid’s Elements*), is still extant, preserved in the Beijing Library. This book aroused interest in China because it was the basis of Western astronomy and therefore offered a new approach to the calendar and to the prediction of eclipses. According to Mikami (1913, p. 114), the Western methods made a correct prediction of a solar eclipse in 1629, which traditional Chinese methods got wrong. It was this accurate prediction that attracted the attention of Chinese mathematicians to Euclid’s book, rather than the elaborate logical structure which is its most prominent distinguishing characteristic. Martzloff (1993) has studied a commented (1700) edition of Euclid by the mathematician Du Zhigeng and has noted that it was considerably abridged, omitting many proofs of propositions that are visually or topologically obvious. As Martzloff says, although Du Zhigeng retained the logical form of Euclid, that is, the definitions, axioms, postulates, and propositions, he neglected proofs, either omitting them entirely or giving only a fraction of a proof, “a fraction not necessarily containing the part of the Euclidean argument relative to a given proposition and

devoted to the mathematical proof in the proper sense of the term." Du Zhigeng also attempted to synthesize the traditional Chinese classics, such as the *Jiu Zhang Suanshu* and the *Suan Fa Tong Zong*, with works imported from Europe, such as Archimedes' treatise on the measurement of the circle. Thus in China, Western mathematics supplemented, but did not replace, the mathematics that already existed.

The first Manchu Emperor Kang Xi (1654–1722) was fascinated by science and insisted on being taught by two French Jesuits, Jean-François Gerbillon (1654–1707) and Joachim Bouvet (1656–1730), who were in China in the late 1680s. This was the time of the Sun King, Louis XIV, who was vying with Spain and Portugal for influence in the Orient. The two Jesuits were required to be at the palace from before dawn until long after sunset and to give lessons to the Emperor for four hours in the middle of each day (Li and Du, 1987, pp. 217–218).

The encounter with the West came at a time when mathematics was undergoing an amazing efflorescence in Europe. The first books on the use of Hindu–Arabic numerals for computation had appeared some centuries before, and now trigonometry, logarithms, analytic geometry, and calculus were all being developed at a rapid pace. These new developments took a firmer hold in China than the ancient Greek mathematics of Euclid and Archimedes. Jami (1988) reports on an eighteenth-century work by Ming Antu (d. 1765) deriving power-series expansions for certain trigonometric functions. She notes that even though the proofs of these expansions would not be regarded as conclusive today, the greatest of the eighteenth-century Chinese mathematicians, Wang Lai (1768–1813), professed himself satisfied with them.⁹ Thus, she concludes, there was a difference between the reception of Euclid in China and the reception of the more computational modern mathematics. The Chinese took Euclid's treatise on its own terms and attempted to fit it into their own conception of mathematics; but they reinterpreted contemporary mathematics completely, since it came to them in small pieces devoid of context (Jami, 1988, p. 327).

Given the increasing contacts between East and West in the nineteenth century, some merging of ideas was inevitable. During the 1850s the mathematician Li Shanlan (1811–1882), described by Martzloff (1982) as "one of the last representatives of Chinese traditional mathematics," translated a number of contemporary works into Chinese, including an 1851 calculus textbook of the American astronomer-mathematician Elias Loomis (1811–1889) and an algebra text by Augustus de Morgan (1806–1871). Li Shanlan had a power over formulas that reminds one in many ways of the twentieth-century Indian genius Srinivasa Ramanujan. One of his combinatorial formulas, stated without proof in 1867, was finally proved through the ingenuity of the prominent Hungarian mathematician Paul Turán (1910–1976). By the early twentieth century Chinese mathematical schools had marked out their own territory, specializing in standard areas of mathematics such as analytic function theory. Despite the difficulties of war, revolution, and a period of isolation during the 1960s, transmission of mathematical literature between China and the West continued and greatly expanded through exchanges of students and faculty from the 1980s onward. Kazdan (1986) gives an interesting snapshot of the situation in China at the beginning of this period of expansion.

⁹ European mathematicians of the time also used methods that would not be considered completely rigorous today, and their arguments have some resemblance to those reported by Jami.

4. Ancient Egypt

Although mathematics has been practiced in Egypt continuously starting at least 4000 years ago, it merged with Greek mathematics during the Hellenistic period that began at the end of the fourth century BCE, and it formed part of the larger Muslim culture centered in Baghdad starting about 1200 years ago. What we shall call Egyptian mathematics in this section had a beginning and an end. It began with hieroglyphic inscriptions containing numbers and dating to the third millennium BCE and ended in the time of Euclid, around 300 BCE. The city of Alexandria in the Nile delta was the main school of mathematics in the Hellenistic world, and many of the most prominent mathematicians who wrote in Greek studied there.

The great architectural monuments of ancient Egypt are covered with hieroglyphs, some of which contain numbers. In fact, the ceremonial mace of the founder of the first dynasty contains records that mention oxen, goats, and prisoners and contain hieroglyphic symbols for the numbers 10,000, 100,000, and 1,000,000. These hieroglyphs, although suitable for ceremonial recording of numbers, were not well adapted for writing on papyrus or leather. The language of the earliest written documents that have been preserved to the present time is a cursive known as *hieratic*.

The most detailed information about Egyptian mathematics comes from a single document written in the hieratic script on papyrus around 1650 BCE and preserved in the dry Egyptian climate. This document is known properly as the Ahmose Papyrus, after its writer, but also as the Rhind Papyrus after the British lawyer Alexander Rhind (1833–1863), who went to Egypt for his health and became an Egyptologist. Rhind purchased the papyrus in Luxor, Egypt, in 1857. Parts of the original document have been lost, but a section consisting of 14 sheets glued end to end to form a continuous roll $3\frac{1}{2}$ feet wide and 17 feet long remains. Part of it is on public display in the British Museum, where it has been since 1865 (see Plate 1). Some missing pieces of this document were discovered in 1922 in the Egyptian collection of the New York Historical Society; these are now housed at the Brooklyn Museum of Art. A slightly earlier mathematical papyrus, now in the Moscow Museum of Fine Arts, consists of sheets about one-fourth the size of the Ahmose Papyrus. This papyrus was purchased by V. S. Golenishchev (1856–1947) in 1893 and donated to the museum in 1912. A third document, a leather roll purchased along with the Ahmose Papyrus, was not unrolled for 60 years after it reached the British Museum because the curators feared it would disintegrate if unrolled. It was some time before suitable techniques were invented for softening the leather, and the document was unrolled in 1927. The contents turned out to be a collection of 26 sums of unit fractions, from which historians were able to gain insight into Egyptian methods of calculation. A fourth set of documents, known as the Reisner Papyri after the American archaeologist George Andrew Reisner (1867–1942), who purchased them in 1904, consists of four rolls of records from dockyard workshops, apparently from the reign of Senusret I (1971–1926 BCE). They are now in the Boston Museum of Fine Arts. Another document, the Akhmim Wooden Tablet, is housed in the Egyptian Museum in Cairo. These documents show the practical application of Egyptian mathematics in construction and commerce.

We are fortunate to be able to date the Ahmose Papyrus with such precision. The author himself gives us his name and tells us that he is writing in the fourth month of the flood season of the thirty-third year of the reign of Pharaoh A-user-re

(Apepi I). From this information Egyptologists arrived at a date of around 1650 BCE for this papyrus. Ahmose tells us, however, that he is merely copying work written down in the reign of Pharaoh Ny-maat-re, also known as Amenemhet III (1842–1797 BCE), the sixth pharaoh of the Twelfth Dynasty. From that information it follows that the mathematical knowledge contained in the papyrus is nearly 4000 years old.

What do these documents tell us about the practice of mathematics in ancient Egypt? Ahmose begins his work by describing it as a “correct method of reckoning, for grasping the meaning of things, and knowing everything that is, obscurities... and all secrets.”¹⁰ The author seems to value mathematics because of its explanatory power, but that explanatory power was essentially practical. The problems that are solved bear a very strong resemblance to those in other treatises such as the *Jiu Zhang Suanshu*.

The Akhmim Wooden Tablet contains several ways of expressing reciprocals of integers based on dividing unity ($64/64$) by these integers. According to Milo Gardner,¹¹ the significance of the number 64 is that it is the number of *ro* in a *hekat* of grain. This origin for the numbers makes sense and gives a solid practical origin for Egyptian arithmetic.

5. Mesopotamia

Some quite sophisticated mathematics was developed four millennia ago in the portion of the Middle East now known as Iraq and Turkey. Unfortunately, this knowledge was preserved on small clay tablets, and nothing like a systematic treatise contemporary with this early mathematics exists. Scholars have had to piece together a mosaic picture of this mathematics from a few hundred clay tablets that show how to solve particular problems. In contrast to Egypt, which had a fairly stable culture throughout many millennia, the region known as Mesopotamia (Greek for “between the rivers”) was the home of many civilizations. The name of the region derives from the two rivers, the Euphrates and the Tigris, that flow from the mountainous regions around the Mediterranean, Black, and Caspian seas into the Persian Gulf. In ancient times this region was a very fertile floodplain, although it suffered from an unpredictable climate. It was invaded and conquered many times, and the successive dynasties spoke and wrote in many different languages. The convention of referring to all the mathematical texts that come from this area between 2500 and 300 BCE as “Babylonian” gives undue credit to a single one of the many dynasties that ruled over this region. The cuneiform script is used for writing several different languages. The tablets themselves date to the period from 2000 to about 300 BCE.

Of the many thousands of cuneiform texts scattered through museums around the world, a few hundred have been found to be mathematical in content. Deciphering them has not been an easy task, although the work was made simpler by multilingual tablets that were created because the cuneiform writers themselves had need to know what had been written in earlier languages. It was not until 1854 that enough tablets had been deciphered to reveal the system of computation used, and not until the early twentieth century were significant numbers of mathematical

¹⁰ This is the translation given by Robins and Shute (1987, p. 11). Chace (1927, p. 49) gives the translation as “the entrance into the knowledge of all existing things and all secrets.”

¹¹ See <http://mathworld.wolfram.com/AkhmimWoodenTablet.html>.

texts deciphered and analyzed. The most complete analysis of these is the 1935 two-volume work by Otto Neugebauer (1899–1992), *Mathematische Keilschrifttexte*, recently republished by Springer-Verlag. A more up-to-date study has been published by the Oxford scholar Eleanor Robson (1999).



Cuneiform tablet BMP 15 285. © The British Museum.

Since our present concern is to introduce authors and their works and discuss motivation, there is little more to say about the cuneiform tablets at this point, except to speculate on the uses for these tablets. Some of the tablets that have been discussed by historians of mathematics appear to be “classroom materials,” written by teachers as exercises for students. This conclusion is based on the fact that the answers so often “come out even.” As Robson (1995, p. 11, quoted by Melville, 2002, p. 2) states, “Problems were constructed from answers known beforehand.” Melville provides an example of a different kind from tablet 4652 of the Yale Babylonian Collection in which the figures are not adjusted this way, but a

certain technique is presumed. Thus, although there is an unavoidable lack of unity and continuity in the Mesopotamian texts compared with mathematics written on more compact and flexible media, the cuneiform tablets nevertheless contain many problems like those considered in India, China, and Egypt. The applications that these techniques had must be inferred, but we may confidently assume that they were the same everywhere: commerce, government administration, and religious rites, all of which call for counting and measuring objects on Earth and making mathematical observations of the sky in order to keep track of months and years.

6. The Maya

The Maya civilization of southern Mexico and Central America began around 2600 BCE. Its period of greatest material wealth lasted from the third to tenth centuries CE. Archaeologists have found evidence of economic decline from the tenth century onward. This civilization was conquered by Spanish explorers in the sixteenth century, with devastating effect on the ancient culture. Some authorities estimate that as much as 90% of the population may have perished of smallpox. Those who survived were dispersed into the countryside and forbidden to practice their ancient religion. In the 1550s the Franciscan friar Diego de Landa (1524–1579) undertook to destroy all Maya books.¹² Fortunately, three Maya books had already been sent to Europe by earlier colonizers. From these few precious remnants, something can be learned of Maya religion and astronomy, which are their main subjects. The authors remain unknown,¹³ so that the books are named for their present locations: the Dresden Codex, the Madrid Codex, and the Paris Codex. A fourth work, consisting of parts of 11 pages of Venus tables, was recovered in Mexico in 1965. It was shown to the Maya scholar Michael Coe, who published it (Coe, 1973). It is known as the Grolier Codex, after the New York publisher of Coe's book, and it now resides in Mexico City.

6.1. The Dresden Codex. The information on the Dresden Codex given here comes from the following website.

<http://www.tu-dresden.de/slub/proj/maya/maya.html>

An English summary of this information can be found at the following website.

<http://www.tu-dresden.de/slub/proj/maya/mayaeng.html>

The codex that is now in the Sächsische Landesbibliothek of Dresden was purchased in Vienna in 1739 by Johann Christian Goetze (1692–1749), who was at the time director of the royal library at the court of Saxony. It is conjectured that the codex was sent to the Hapsburg Emperor Charles V (1500–1558; he was also King Carlos I of Spain). The codex consists of 74 folios, folded like an accordion, with texts and illustrations in bright colors (see Plate 3). It suffered some damage from the British–American bombing of Dresden during World War II. Fortunately,

¹² Sources differ on the dates of Diego de Landa's life. Ironically, de Landa's own work helped in deciphering the Maya hieroglyphs, which he tried to summarize in a history of events in the Yucatan. He had no quarrel with the Maya language, only with the religious beliefs embodied in its books. According to Sharer (1994, p. 558), de Landa recognized and wrote about the vigesimal place system used by the Maya.

¹³ One can infer something about Maya mathematics from the remains of Maya crafts and architecture. Closs (1992, p. 12) notes that one Maya vase contains a painting with a scribe figure who is apparently female and bears the name "Ah Ts'ib, The Scribe."

considerable work had been done earlier on the codex by another director of the Dresden library, a philologist named Ernst Förstemann (1822–1906), who had 200 copies of it made.

The work consists of eight separate treatises and, according to the experts, shows evidence of having been written by eight different people. Dates conjectured for it vary from the thirteenth to the fifteenth centuries, and it may have been a copy of an earlier document. The first 15 folios are devoted to almanacs and astronomy/astrology, while folios 16–23 are devoted to the Moon Goddess. Both of these sections are based on a 260-day calendar known as the *Tzolkin* (see Chapter 5). It is believed that these pages were consulted to determine whether the gods were favorably inclined toward proposed undertakings. Folio 24 and folios 46–50 are Venus tables, containing 312 years of records of the appearance of Venus as morning and evening star. Such records help to establish the chronology of Maya history as well as the date of the manuscript itself. The pictures accompanying the text (Plate 3) seem to indicate a belief that Venus exerted an influence on human life. These pages are followed by eclipse tables over the 33-year period from 755 to 788 CE. Folios 25–28 describe new-year ceremonies, and folios 29–45 give agricultural almanacs. Folios 61–73 give correlations of floods and storms with the 260-day calendar in order to predict the end of the next world cycle. Finally, folio 74 describes the coming end of the current world cycle. The Maya apparently believed there had been at least three such cycles before the current one.

The mathematics that can be gleaned from these codices and the steles that remain in Maya territory is restricted to applications to astronomy and the calendar. The arithmetic that is definitely attested by documents is rudimentary. For example, although there is no reason to doubt that the Maya performed multiplication and division, there is no record showing how they did so. Undoubtedly, there was a Maya arithmetic for commerce, but it is very difficult to reconstruct, since no treatises on the subject exist. Thus, our understanding of the achievements of Maya scientists and mathematicians is limited by the absence of sources. The Maya documents that have survived to the present are “all business” and contain no whimsical or pseudo-practical problems of an algebraic type such as can be found in ancient Chinese, Hindu, Mesopotamian, and Egyptian texts.

Questions and problems

2.1. Does mathematics realize Plato’s program of understanding the world by contemplating eternal, unchanging forms that are perceived only by reason, not by the senses?

2.2. To what extent do the points of view expressed by Hamming and Hardy on the value of pure mathematics reflect the nationalities of their authors and the prevailing attitudes in their cultures? Consider that unlike the public radio and television networks in the United States, the CBC in Canada and the BBC in Britain do not spend four weeks a year pleading with their audience to send voluntary donations to keep them on the air. The BBC is publicly funded out of revenues collected by requiring everyone who owns a television set to pay a yearly license fee.

2.3. In an article in the *Review of Modern Physics*, **51**, No. 3 (July 1979), the physicist Norman David Mermin (b. 1935) wrote, “Bridges would not be safer if only people who knew the proper definition of a real number were allowed to design

them" (quoted by Mackay, 1991, p. 172). Granting that at the final point of contact between theory and the physical world, when a human design is to be executed in concrete and steel, every number is only an approximation, is there any value for science and engineering in the concept of an infinitely precise real number? Or is this concept only for idealistic, pure mathematicians? (The problems below may influence your answer.)

2.4. In 1837 and 1839 the crystallographer Auguste Bravais (1811–1863) and his brother Louis (1801–1843) published articles on the growth of plants.¹⁴ In these articles they studied the spiral patterns in which new branches grow out of the limbs of certain trees and classified plants into several categories according to this pattern. For one of these categories they gave the amount of rotation around the limb between successive branches as $137^{\circ} 30' 28''$. Now, one could hardly measure the limb of a tree so precisely. To measure within 10° would require extraordinary precision. To refine such crude measurements by averaging to the claimed precision of $1''$, that is, $1/3600$ of a degree, would require thousands of individual measurements. In fact, the measurements were carried out in a more indirect way, by counting the total number of branches after each full turn of the spiral. Many observations convinced the brothers Bravais that normally there were slightly more than three branches in two turns, slightly less than five in three turns, slightly more than eight in five turns, and slightly less than thirteen in eight turns. For that reason they took the actual amount of revolution between successive branches to be the number we call $1/\Phi = (\sqrt{5} - 1)/2 = \Phi - 1$ of a complete (360°) revolution, since

$$\frac{3}{2} < \frac{8}{5} < \Phi < \frac{13}{8} < \frac{5}{3}.$$

Observe that $360^{\circ} \div \Phi \approx 222.4922359^{\circ} \approx 222^{\circ} 29' 32'' = 360^{\circ} - (137^{\circ} 30' 28'')$. Was there scientific value in making use of this *real* (infinitely precise) number Φ even though no actual plant grows exactly according to this rule?

2.5. Plate 4 shows a branch of a flowering crab apple tree from the author's garden with the twigs cut off and the points from which they grew marked by pushpins. The "zeroth" pin at the left is white. After that, the sequence of colors is red, blue, yellow, green, pink, clear, so that the red pins correspond to 1, 7, and 13, the blue to 2 and 8, the yellow to 3 and 9, the green to 4 and 10, the pink to 5 and 11, and the clear to 6 and 12. (The green pin corresponding to 4 and part of the clear pin corresponding to 12 are underneath the branch and cannot be seen in the picture.) Observe that when these pins are joined by string, the string follows a helical path of nearly constant slope along the branch. Which pins fall nearest to the intersection of this helical path with the meridian line marked along the length of the branch? How many turns of the spiral correspond to these numbers of twigs? On that basis, what is a good approximation to the number of twigs per turn? Between which pin numbers do the intersections between the spiral and the meridian line fall? For example, the fourth intersection is between pins 6 and 7, indicating that the average number of pins per turn up to that point is between

¹⁴ See the article by I. Adler, D. Barabe, and R. V. Jean, "A history of the study of phyllotaxis," *Annals of Botany*, **80** (1997), 231–244, especially p. 234. The articles by Auguste and Louis Bravais are "Essai sur la disposition générale des feuilles curvisériées," *Annales des sciences naturelles*, **7** (1837), 42–110, and "Essai sur la disposition générale des feuilles rectisériées," *Congrès scientifique de France*, **6** (1839), 278–330.

$\frac{6}{4} = 1.5$ and $\frac{7}{4} = 1.75$. Get upper and lower estimates in this way for all numbers of turns from 1 to 8. What are the narrowest upper and lower bounds you can place on the number of pins per turn in this way?

2.6. Suppose that the pins in Plate 4 had been joined by a curve winding in the opposite direction. How would the numbers of turns of the spiral and the number of pins joined compare? What change would occur in the slope of the spiral?

2.7. With which of the two groups of people mentioned by Plato do you find yourself more in sympathy: the “practical” people, who object to being taxed to support abstract speculation, or the “idealists,” who regard abstract speculation as having value to society?

2.8. The division between the practical and the ideal in mathematics finds an interesting reflection in the interpretation of what is meant by solving an equation. Everybody agrees that the problem is to find a number satisfying the equation, but interpretations of “finding a number” differ. Inspired by Greek geometric methods, the Muslim and European algebraists looked for algorithms to invert the operations that defined the polynomial whose roots were to be found. Their object was to generate a sequence of arithmetic operations and root extractions that could be applied to the coefficients in order to exhibit the roots. The Chinese, in contrast, looked for numerical processes to approximate the roots with arbitrary accuracy. What advantages and disadvantages do you see in each of these approaches? What would be a good synthesis of the two methods?

2.9. When a mathematical document such as an early treatise or cuneiform tablet contains problems whose answers “come out even,” should one suspect or conclude that it was a teaching device—either a set of problems with simplified data to build students’ confidence or a manual for teachers showing how to construct such problems?

2.10. From what is known of the Maya codices, is it likely that they were textbooks intended for teaching purposes, like many of the cuneiform tablets and the early treatises from India, China, and Egypt?

2.11. Why was the Chinese encounter with the Jesuits so different from the Maya encounter with the Franciscans? What differences were there in the two situations, and what conditions account for these differences? Was it merely a matter of the degree of zeal that inspired Diego de Landa and Matteo Ricci, or were there institutional or national differences between the two as well? How much difference did the relative strength of the Chinese and the Maya make?

CHAPTER 3

Mathematical Cultures II

Many cultures borrow from others but add their own ideas to what they borrow and make it into something richer than the pure item would have been. The Greeks, for example, never concealed their admiration for the Egyptians or the debt that they owed to them, yet the mathematics that they passed on to the world was vastly different from what they learned in Egypt. To be sure, much of it was created in Egypt, even though it was written in Greek. The Muslim culture that flourished from 800 to 1500 CE learned something from the Greeks and Hindus, but also made many innovations in algebra, geometry, and number theory. The Western Europeans are still another example. Having learned about algebra and number theory from the Byzantine Empire and the Muslims, they went on to produce such a huge quantity of first-rate mathematics that for a long time European scholars were tempted to think of the rest of the world as merely a footnote to their own work. For example, in his history of western philosophy (1945), the British philosopher Bertrand Russell wrote (p. xvi), "In the Eastern Empire, Greek civilization, in a desiccated form, survived, as in a museum, till the fall of Constantinople in 1453, but nothing of importance to the world came out of Constantinople except an artistic tradition and Justinian's Codes of Roman law." He wrote further (p. 427), "Arabic philosophy is not important as original thought. Men like Avicenna and Averroes are essentially commentators." Yet Russell was not *consciously* a chauvinist. In the same book (p. 400), he wrote, "I think that, if we are to feel at home in the world after the present war [World War II], we shall have to admit Asia to equality in our thoughts, not only politically, but culturally."

Until very recently, many textbooks regarded as authoritative were written from this "it all started with the Greeks" point of view. As Chapter 2 has shown, however, there was mathematics before the Greeks, and the Greeks learned it before they began making their own remarkable innovations in it.

1. Greek and Roman mathematics

The Greeks of the Hellenic period (to the end of the fourth century BCE) traced the origins of their mathematical knowledge to Egypt and the Middle East. This knowledge probably came in "applied" form in connection with commerce and astronomy/astrology. The evidence of Mesopotamian numerical methods shows up most clearly in the later Hellenistic work on astronomy by Hipparchus (second century BCE) and Ptolemy (second century CE). Earlier astronomical models by Eudoxus (fourth century BCE) and Apollonius (third century BCE) were more geometrical. Jones (1991, p. 445) notes that "the astronomy that the Hellenistic Greeks received from the hands of the Babylonians was by then more a skill than a science: the quality of the predictions was proverbial, but in all likelihood the practitioners knew little or nothing of the origins of their schemes in theory and

observations.” Among the techniques transmitted to the Greeks and ultimately to the modern world was the convention of dividing a circle into 360 equal parts (degrees). Greek astronomers divided the radius into 60 equal parts so that the units of length on the radius and on the circle were very nearly equal.

The amount that the Greeks learned from Egypt is the subject of controversy. Many scholars who have read the surviving mathematical texts from papyri have concluded that Egyptian methods of computing were too cumbersome for application to the complicated measurements of astronomers. Yet both Plato and Aristotle speak approvingly of Egyptian computational methods and the ways in which they were taught. As for geometry, it is generally acknowledged that the Egyptian insight was extraordinary; the Egyptians knew how to find the volume of a pyramid, for example. They even found the area of a hemisphere, the only case known before Archimedes in which the area of a curved surface is found.¹ The case for advanced Egyptian mathematics is argued in some detail by Bernal (1992), who asserts that Ptolemy himself was an Egyptian. The argument is difficult to settle, since little is known of Ptolemy personally; for us, he is simply the author of certain works on physics and astronomy.

Because of their extensive commerce, with its need for counting, measuring, navigation, and an accurate calendar, the Ionian Greek colonies such as Miletus on the coast of Asia Minor and Samos in the Aegean Sea provided a very favorable environment for the development of mathematics, and it was there, with the philosophers Thales of Miletus (ca. 624–547 BCE) and Pythagoras of Samos (ca. 570–475 BCE), that Greek mathematics began.

1.1. Sources. Since the material on which the Greeks wrote was not durable, all the original manuscripts have been lost except for a few ostraca (shells) found in Egypt. We are dependent on copyists for preserving the information in early Greek works, since few manuscripts that still exist were written more than 1000 years ago. We are further indebted to the many commentators who wrote summary histories of philosophy, including mathematics, for the little that we know about the works that have not been preserved and their authors. The most prominent among these commentators are listed below. They will be mentioned many times in the chapters that follow.

Marcus Vitruvius (first century BCE) was a Roman architect who wrote an extremely influential treatise on architecture in 10 books. He is regarded as a rather unreliable source for information about mathematics, however.

Plutarch (45–120 CE) was a pagan author, apparently one of the best educated people of his time, who wrote on many subjects. He is best remembered as the author of the *Parallel Lives of the Greeks and Romans*, in which he compares famous Greeks with eminent Romans who engaged in the same occupation, such as the orators Demosthenes and Cicero.² Plutarch is important to the history of mathematics for what he reports on natural philosophers such as Thales.

¹ Some authors claim that the surface in question was actually half of the lateral surface of a cylinder, but the words used seem more consistent with a hemisphere. In either case it was a curved surface.

² Shakespeare relied on Plutarch's account of the life of Julius Caesar, even describing the miraculous omens that Plutarch reported as having occurred just before Caesar's death.

Theon of Smyrna (ca. 100 CE) was the author of an introduction to mathematics written as background for reading Plato, a copy of which still exists. It contains many quotations from earlier authors.

Diogenes Laertius (third century CE) wrote a comprehensive history of philosophy, *Lives of Eminent Philosophers*, which contains summaries of many earlier works and gives details of the lives and work of many of the pre-Socratic philosophers. He appears to be the source of the misnomer "Pythagorean theorem" that has come down to us (see Zhmud, 1989, p. 257).

Iamblichus (285–330 CE) was the author of many treatises, including 10 books on the Pythagoreans, five of which have been preserved.

Pappus (ca. 300 CE) wrote many books on geometry, including a comprehensive treatise of eight mathematical books. He is immortalized in calculus books for his theorem on the volume of a solid of revolution. Besides being a first-rate geometer in his own right, he wrote commentaries on the *Almagest* of Ptolemy and the tenth book of Euclid's *Elements*.

Proclus (412–485 CE) is the author of a commentary on the first book of Euclid, in which he quoted a long passage from a history of mathematics, now lost, by Eudemus, a pupil of Aristotle.

Simplicius (500–549 CE) was a commentator on philosophy. His works contain many quotations from the pre-Socratic philosophers.

Eutocius (ca. 700 CE) was a mathematician who lived in the port city of Askalon in Palestine and wrote an extensive commentary on the works of Archimedes.

Most of these commentators wrote in Greek. Knowledge of Greek sank to a very low level in western Europe as a result of the upheavals of the fifth century. Although learning was preserved by the Church and all of the New Testament was written in Greek, a Latin translation (the Vulgate) was made by Jerome in the fifth century. From that time on, Greek documents were preserved mostly in the Eastern (Byzantine) Empire. After the Muslim conquest of North Africa and Spain in the eighth century, some Greek documents were translated into Arabic and circulated in Spain and the Middle East. From the eleventh century on, as secular learning began to revive in the West, scholars from northern Europe made journeys to these centers and to Constantinople, copied out manuscripts, translated them from Arabic and Greek into Latin, and tried to piece together some long-forgotten parts of ancient learning.

1.2. General features of Greek mathematics. Greek mathematics—that is, mathematics written in ancient Greek—is exceedingly rich in authors and works. Its most unusual feature, compared with what went before, is its formalism. Mathematics is developed systematically from definitions and axioms, general theorems are stated, and proofs are given. This formalism is the outcome of the entanglement of mathematics with Greek philosophy. It became a model to be imitated in many later scientific treatises, such as Newton's *Philosophiæ naturalis principia mathematica*. Of course, Greek mathematics did not arise in the finished form found in the treatises. Tradition credits Thales only with knowing four geometric propositions. By the time of Pythagoras, much more was known. The crucial formative period was the first half of the fourth century BCE, when Plato's Academy flourished. Plato himself was interested in mathematics because he hoped for a sort of "theory of everything," based on fundamental concepts perceived by the mind.

Plato is famous for his theory of ideas, which had both metaphysical and epistemological aspects. The metaphysical aspect was a response to two of his predecessors, Heraclitus of Ephesus (ca. 535–475 BCE), who asserted that everything is in constant flux, and Parmenides (born around 515 BCE), who asserted that knowledge is possible only in regard to things that do not change. One can see the obvious implication: Everything changes (Heraclitus). Knowledge is possible only about things that do not change (Parmenides). *Therefore...* To avoid the implication that no knowledge is possible, Plato restricted the meaning of Heraclitus' "everything" to objects of sense and invented eternal, unchanging Forms that could be objects of knowledge.

The epistemological aspect of Plato's philosophy involves universal propositions, statements such as "Lions are carnivorous" (our example, not Plato's), meaning "All lions are carnivorous." This sentence is grammatically inconsistent with its meaning, in that the grammatical subject is the set of all lions, while the assertion is not about this set but about its individual members. It asserts that each of them is a carnivore, and therein lies the epistemological problem. What is the real subject of this sentence? It is not any particular lion. Plato tried to solve this problem by inventing the Form or Idea of a lion and saying that the sentence really asserts a relation perceived in the mind between the Form of a lion and the Form of a carnivore. Mathematics, because it dealt with objects and relations perceived by the mind, appeared to Plato to be the bridge between the world of sense and the world of Forms. Nevertheless, mathematical objects were not the same thing as the Forms. Each Form, Plato claimed, was unique. Otherwise, the interpretation of sentences by use of Forms would be ambiguous. But mathematical objects such as lines are not unique. There must be at least three lines, for example, in order for a triangle to exist. Hence, as a sort of hybrid of sense experience and pure mental creation, mathematical objects offered a way for the human soul to ascend to the height of understanding, by perceiving the Forms themselves. Incorporating mathematics into education so as to realize this program was Plato's goal, and his pupils studied mathematics in order to achieve it. Although the philosophical goal was not reached, the effort expended on mathematics was not wasted; certain geometric problems were solved by people associated with Plato, providing the foundation of Euclid's famous work, known as the *Elements*.

Within half a century of Plato's death, Euclid was writing that treatise, which is quite free of all the metaphysical accoutrements that Plato's pupils had experimented with. However, later neo-Platonic philosophers such as Proclus attempted to reintroduce philosophical ideas into their commentary on Euclid's work. The historian and mathematician Otto Neugebauer (1975, p. 572) described the philosophical aspects of Proclus' introduction as "gibberish," and expressed relief that scientific methodology survived despite the prevalent dogmatic philosophy.

According to Diels (1951, 44A5), Plato met the Pythagorean Philolaus in Sicily in 390. In any case, Plato must certainly have known the work of Philolaus, since in the *Phaedo* Socrates says that both Cebes and Simmias are familiar with the work of Philolaus and implies that he himself knows of it at second hand. It seems likely, then, that Plato's interest in mathematics began some time after the death of Socrates and continued for the rest of his life, that mathematics played an important role in the curriculum of his Academy and in the research conducted there, and that Plato himself played a leading role in directing that research. We do not, however, have any theorems that can with confidence be attributed to Plato

himself. Lasserre (1964, p. 17) believes that the most important mathematical work at the Academy was done between 375 and 350 BCE.

Socrates explained that arithmetic was needed both to serve the eye of the soul and as a practical instrument in planning civic projects and military campaigns:

The kind of knowledge we are seeking seems to be as follows. It is necessary for a military officer to learn (*matheîn*) these things for the purpose of proper troop deployment, and the philosopher must have risen above change, in order to grasp the essence of things, or else never become skilled in calculation (*logistikô*).

Plato, through Socrates, complains of the lack of a government subsidy for geometry. In his day solid geometry was underdeveloped in comparison with plane geometry, and Socrates gave what he thought were the reasons for its backwardness:

First, no government holds [the unsolved problems in solid geometry] in honor; and they are researched in a desultory way, being difficult. Second, those who are doing the research need a mentor, without which they will never discover anything. But in the first place, to become a mentor is difficult; and in the second place, after one became a mentor, as things are just now, the arrogant people doing this research would never listen to him. But if the entire state were to act in concert in conducting this research with respect, the researchers would pay heed, and by their combined intensive work the answers would become clear.

Plato himself was among that group of people mentioned in Chapter 2, for whom the “eye of the soul” was sufficient justification for intellectual activity. He seems to have had a rather dim view of the second group, the practical-minded people. In his long dialogue *The Laws*, one of the speakers, an Athenian, rants about the shameful Greek ignorance of incommensurables, surely a topic of limited application in the lives of most people.

1.3. Works and authors. Books on mathematics written in Greek begin appearing early in Hellenistic times (third century BCE) and continue in a steady stream for hundreds of years. We list here only a few of the most outstanding authors.

Euclid. This author lived and worked in Alexandria, having been invited by Ptolemy Soter (Ptolemy I) shortly after the city was founded. Essentially nothing is known of his life beyond that fact, but his famous treatise on the basics of geometry (the *Elements*) has become a classic known all over the world. Several of his minor works—the *Optics*, the *Data*, and the *Phænomena*—also have been preserved. Unlike Aryabhata and Bhaskara, Euclid did not provide any preface to tell us why he wrote his treatise. We do, however, know enough of the Pythagorean philosophy to understand why they developed geometry and number theory to the extent that they did, and it is safe to conclude that this kind of work was considered valuable because it appealed to the intellect of those who could understand it.

Archimedes. Much more is known of Archimedes (ca. 287–212 BCE). About 10 of his works have been preserved, including the prefaces that he wrote in the form of “cover letters” to the people who received the works. Here is one such letter, which accompanied a report of what may well be regarded as his most profound achievement—proving that the area of a sphere is four times the area of its equatorial circle.

On a former occasion I sent you the investigations which I had up to that time completed, including the proofs, showing that any segment bounded by a straight line and a section of a right-angled cone [parabola] is four-thirds of the triangle which has the same base with the segment and equal height. Since then certain theorems not hitherto demonstrated have occurred to me, and I have worked out the proofs of them. They are these: first, that the surface of any sphere is four times its greatest circle. . . For, though these properties also were naturally inherent in the figures all along, yet they were in fact unknown to all the many able geometers who lived before Eudoxus, and had not been observed by anyone. Now, however, it will be open to those who possess the requisite ability to examine these discoveries of mine. [Heath, 1897, Dover edition, pp. 1–2]

As this letter shows, mathematics was a “going concern” by Archimedes’ time, and a community of mathematicians existed. Archimedes is known to have studied in Alexandria. He perished when his native city of Syracuse was taken by the Romans during the Second Punic War. Some of Archimedes’ letters, like the one quoted above, give us a glimpse of mathematical life during his time. Despite being widely separated, the mathematicians of the time sent one another challenges and communicated their achievements.

Apollonius. Apollonius, about one generation younger than Archimedes, was a native of what is now Turkey. He studied in Alexandria somewhat after the time of Euclid and is also said to have taught there. He eventually settled in Pergamum (now Bergama in Turkey). He is the author of eight books on conic sections, four of which survive in Greek and three others in an Arabic translation. We know that there were originally eight books because commentators, especially Pappus, described the work and told how many propositions were in each book.

In his prefaces Apollonius implies that geometry was simply part of what an educated person would know, and that such people were as fascinated with it in his time as they are today about the latest scientific achievements. Among other things, he said the following.

During the time I spent with you at Pergamum I observed your eagerness to become acquainted with my work in conics. [Book I]

I undertook the investigation of this subject at the request of Naurates the geometer, at the time when he came to Alexandria and stayed with me, and, when I had worked it out in eight books, I gave them to him at once, too hurriedly, because he was on the point of sailing; they had therefore not been thoroughly revised,

indeed I had put down everything just as it occurred to me, postponing revision until the end. [Book II]

Ptolemy. Claudius Ptolemy was primarily an astronomer and physicist, although these subjects were hardly distinct from mathematics in his time. He lived in Alexandria during the second century, as is known from the astronomical observations that he made between 127 and 141 CE. He created an intricate and workable Earth-centered mathematical system of explaining the motion of the planets and systematized it in a treatise known as the *Syntaxis*, which, like Euclid's, consisted of 13 books. Also like Euclid's treatise, Ptolemy's *Syntaxis* became a classic reference and was used for well over a thousand years as the definitive work on mathematical astronomy. It soon became known as the "greatest" work (*megistos* in Greek) and when translated into Arabic became *al-megista* or the *Almagest*, as we know it today.

Diophantus. Little is known about this author of a remarkable treatise on what we now call algebra and number theory. He probably lived in the third century CE, although some experts believe he lived earlier than that. His treatise is of no practical value in science or commerce, but its problems inspired number theorists during the seventeenth century and led to the long-standing conjecture known as Fermat's last theorem. The 1968 discovery of what may be four books from this treatise that were long considered lost was the subject of a debate among the experts, some of whom believed the books might be commentaries, perhaps written by the late fourth-century commentator Hypatia. If so, they would be the only work by Hypatia still in existence.

Pappus. Pappus, who is known to have observed a solar eclipse in Alexandria in 320 CE, was the most original and creative of the later commentators on Greek geometry and arithmetic. His *Synagōgē* (*Collection*) consists of eight books of insightful theorems on arithmetic and geometry, as well as commentary on the works of other authors. In some cases where works of Euclid, Apollonius, and others have been lost, this commentary tells something about these works. Pappus usually assumes that the reader is interested in what he has to say, but sometimes he gives in addition a practical justification for his study, as in Book 8:

The science of mechanics, my dear Hermodorus, has many important uses in practical life, and is held by philosophers to be worthy of the highest esteem, and is zealously studied by mathematicians, because it takes almost first place in dealing with the nature of the material elements of the universe. [Thomas, 1941, p. 615]

As a commentator, Pappus was highly original, and the later commentators Theon of Alexandria (late fourth century) and his daughter Hypatia (ca. 370–415) produced respectable work, including a standard edition of Euclid's *Elements*. Several of Theon's commentaries still exist, but nothing authored by Hypatia has been preserved, unless the books of Diophantus mentioned above were written by her. Very little of value can be found in Greek mathematics after the fourth century. As Gow (1884, p. 308) says:

The *Collection* of Pappus is not cited by any of his successors, and none of them attempted to make the slightest use of the proofs and *aperçus* in which the book abounds... His work is only the

last convulsive effort of Greek geometry which was now nearly dead and was never effectually revived.

Greek mathematics held on longer in the Byzantine Empire than in Western Europe. Although Theon of Alexandria had found it necessary to water down the more difficult parts of Greek geometry for the sake of his weak students, the degeneration in Latin works was even greater. The philosopher Boethius (480–524) wrote Latin translations of many classical Greek works of mathematics and philosophy. His works on mathematics were translations based on Nicomachus and Euclid. Boethius' translation of Euclid has been lost. However, it is believed to be the basis of many other medieval manuscripts, some of which use his name. These are referred to as "Boethius" or pseudo-Boethius. The works of Boethius fit into the classical quadrivium of arithmetic, geometry, music, and astronomy.

Politically and militarily, the fifth century was full of disasters in Italy, and some of the best minds of the time turned from public affairs to theological questions. For many of these thinkers mathematics came to be valued only to the extent that it could inspire religious feelings. The pseudo-Boethius gives a good example of this point of view. He writes:³

The utility of geometry is threefold: for work, for health, and for the soul. For work, as in the case of a mechanic or architect; for health, as in the case of the physician; for the soul, as in the case of the philosopher. If we pursue this art with a calm mind and diligence, it is clear in advance that it will illuminate our senses with great clarity and, more than that, will show what it means to subordinate the heavens to the soul, to make accessible all the supernal mechanism that cannot be investigated by reason in any other way and through the sublimity of the mind beholding it, also to integrate and recognize the Creator of the world, who veiled so many deep secrets.

In the tenth century, Gerbert of Aurillac (940–1003), who became Pope Sylvester II in 999, wrote a treatise on geometry based on Boethius. His reasons for studying geometry were similar:

Indeed the utility of this discipline to all lovers of wisdom is the greatest possible. For it leads to vigorous exercises of the soul, and the most subtle demands on the intuition, and to many certain inquiries by true reasoning, in which wonderful and unexpected and joyful things are revealed to many along with the wonderful vigor of nature, and to contemplating, admiring, and praising the power and ineffable wisdom of the Creator who apportioned all things according to number and measure and weight; it is replete with subtle speculations.

These uses of geometry were expressed in the last Canto of Dante's *Divine Comedy*, which describes the poet's vision of heaven:

³ This quotation and the next can be read online at <http://pld.chadwyck.com>, a commercial website. This passage is from Vol. 63; the next is from Vol. 139. Both can be reached by searching under "geometria" as title.

Like the geometer who applies all his powers
To measure the circle, but does not find
By thinking the principle he needs,
Such was I, in this new vista.
I wished to see how the image came together
With the circle and how it could be divined there.
But my own wings could not have made the flight
Had not my mind been struck
By a flash in which his will came to me.
In this lofty vision I could do nothing.
But now turning my desire and will,
Like a wheel that is uniformly moved,
Was the love that moves the sun and the other stars.



The quadrivium, from Boethius' *Arithmetic*. From left to right: Music holding an instrument, Arithmetic doing a finger computation, Geometry studying a set of figures, Astrology holding a set of charts for horoscopes. © Foto Marburg/Art Resource.

The Byzantine Empire and modern Greece. Mathematics continued in the Byzantine Empire until the Turks conquered Constantinople in 1453. Of several figures who contributed to it, the one most worthy of mention is the monk Maximus Planudes (ca. 1260–1310), who is best known for the literature that he preserved (including Aesop's *Fables*). Planudes wrote commentaries on the work of Diophantus and gave an account of the Hindu numerals that was one of the sources from which these numerals eventually came down to us (Heath, 1921, pp. 546–547).

The mainland of Greece was partitioned and disputed among various groups for centuries: the Latin West, the Byzantine Empire, the Ottoman Empire, the Venetians, and the Normans invaded or ruled over parts of it. In the fourteenth century it became a part of the Ottoman Empire, from which it gained independence only in the 1820s and 1830s. Even before independence, however, Greek scholars, inspired by the great progress in Europe, were laying the foundations of a modern mathematical school (see Phili, 1997).

2. Japan

Both Korea and Japan adopted the Chinese system of writing their languages. The Chinese language was the source of a huge amount of technical vocabulary in Korea and Japan over many centuries, and even in recent times in Viet Nam (Koblitz, 1990, p. 26). The establishment of Buddhism in Japan in the sixth century increased the rate of cultural importation from China and even from India.⁴

The influence of Chinese mathematics on both Korea and Japan was considerable. The courses of university instruction in this subject in both countries were based on reading (in the original Chinese language) the Chinese classics we discussed in Chapter 2. In relation to Japan the Koreans played a role as transmitters, passing on Chinese learning and inventions. This transmission process began in 553–554 when two Korean scholars, Wang Lian-tung and Wang Pu-son, journeyed to Japan. For many centuries both the Koreans and the Japanese worked within the system of Chinese mathematics. The earliest records of new and original work in these countries date from the seventeenth century. By that time mathematical activity was exploding in Europe, and Europeans had begun their long voyages of exploration and colonization. There was only a brief window of time during which indigenous mathematics independent of Western influence could grow up in these countries. The following synopsis is based mostly on the work of Mikami (1913), Smith and Mikami (1914), and Murata (1994). Following the usage of the first two of these sources, all Japanese names are given surname first. A word of caution is needed about the names, however. Most Chinese symbols (*kanji* in Japanese) have at least two readings in Japanese. For example, the symbol read as *chu* in the Japanese word for China (*Chugoku*), is also read as *naka* (meaning *middle*) in the surname Tanaka. These variant readings often cause trouble in names from the past, so that one cannot always be sure how a name was pronounced. As Mikami (1913, p. viii) says, “We read Seki Kōwa, although his personal name Kōwa should have been read Takakazu.” Several examples of such alternate readings will be encountered below. A list of these names and their *kanji* rendering can be found in a paper of Martzloff (1990, p. 373).

⁴ The Japanese word for China—*Chūgoku*—means literally *Midland*, that is, between Japan and India.

2.1. Chinese influence and calculating devices. All the surviving Japanese records date from the time after Japan had adopted the Chinese writing system. Japanese mathematicians were for a time content to read the Chinese classics. In 701 the emperor Monbu established a university system in which the mathematical part of the curriculum consisted of 10 Chinese classics. Some of these are no longer known, but the *Zhou Bi Suan Jing*, *Sun Zi Suan Jing*, *Jiu Zhang Suanshu*, and *Hai Dao Suan Jing* were among them. Japan was disunited for many centuries after this early encounter with Chinese culture, and the mathematics that later grew up was the result of a reintroduction in the sixteen and seventeenth centuries. In this reintroduction, the two most important works were the *Suan Fa Tong Zong* by Cheng Dawei, and the *Suan Shu Chimeng* of Zhu Shijie, both mentioned in Chapter 2. The latter became part of the curriculum in Korea very soon after it was written and was published in Japan in the mid-seventeenth century. The evidence of Chinese influence is unmistakable in the mechanical methods of calculation used for centuries—counting rods, counting boards, and the abacus, which played an especially important role in Japan.

The Koreans adopted the Chinese counting rods and counting boards, which the Japanese subsequently adopted from them. The abacus (*suan pan*) was invented in China, probably in the fourteenth century, when methods of computing with counting rods had become so efficient that the rods themselves were a hindrance to the performance of the computation. From China the invention passed to Korea, where it was known as the *sanbob*. Because it did not prove useful in Korean business, it did not become widespread there. It passed on to Japan, where it is known as the *soroban*, which may be related to the Japanese word for an orderly table (*soroiban*). The Japanese made two important technical improvements in the abacus: (1) they replaced the round beads by beads with sharp edges, which are easier to manipulate; and (2) they eliminated the superfluous second 5-bead on each string.

2.2. Japanese mathematicians and their works. A nineteenth-century Japanese historian reported that the emperor Hideyoshi sent the scholar Mōri Shigeyoshi (Mōri Kambei) to China to learn mathematics. According to the story, the Chinese ignored the emissary because he was not of noble birth. When he returned to Japan and reported this fact, the emperor conferred noble status on him and sent him back. Unfortunately, his second visit to China coincided with Hideyoshi's unsuccessful attempt to invade Korea, which made his emissary unwelcome in China. Mōri Shigeyoshi did not return to Japan until after the death of Hideyoshi, but when he did return (in the early seventeenth century), he brought the abacus with him. Whether this story is true or not, it is a fact that Mōri Shigeyoshi was one of the most influential early Japanese mathematicians. He wrote several treatises, all of which have been lost, but his work led to a great flowering of mathematical activity in seventeenth-century Japan, through the work of his students. This mathematics was known as *wasan*, and written using two Chinese characters. The first is *wa*, a word still used to denote Japanese-style work in arts and crafts, meaning literally *harmony*. The second is *san*, meaning calculation, the same Chinese symbol that represents *suan* in the many Chinese classics mentioned above.⁵ Murata (1994, p. 105) notes that the primary concern in *wasan* was to obtain elegant results, even when those results required very complicated calculations, and that "many

⁵ The modern Japanese word for mathematics is *sūgaku*, meaning literally *number study*.

Wasanists were men of fine arts rather than men of mathematics in the European sense."

According to Murata (1994), the stimulus for the development of *wasan* came largely from the two Chinese classics mentioned above, the 1593 arithmetical treatise *Suan Fa Tong Zong* and the algebraic treatise *Suan Shu Chimeng*. The latter was particularly important, since it came with no explanatory notes and a rebellion in China had made communication with Chinese scholars difficult. By the time this treatise was understood, the Japanese mathematicians had progressed beyond its contents.

Sangaku. The shoguns of the Tokugawa family (1600–1868) concentrated their foreign policy on relations with China and held Western visitors at arms length, with the result that Japan was nearly closed to the Western world for 250 years. During this time a fascinating form of mathematics known as *sangaku* (mathematical study, the "study" being a physical plaque) arose, involving the posting of mathematical plaques at sacred shrines (see Plate 2). These problems are discussed in detail in the book of Fukagawa and Pedoe (1989).

Yoshida Koyu. Mōri Shigeyoshi trained three outstanding students during his lifetime, of whom we shall discuss only the first. This student was Yoshida Koyu (Yoshida Mitsuyoshi, 1598–1672). Being handicapped in his studies at first by his ignorance of Chinese, Yoshida Koyu devoted extra effort to this language in order to read the *Suan Fa Tong Zong*. Having read this book, Yoshida Koyu made rapid progress in mathematics and soon excelled even Mōri Shigeyoshi himself. Eventually, he was called to the court of a nobleman as a tutor in mathematics. In 1627 Yoshida Koyu wrote a textbook in Japanese, the *Jinkō-ki* (*Treatise on Large and Small Numbers*), based on the *Suan Fa Tong Zong*. This work helped to popularize the abacus (soroban) in Japan. It concluded with a list of challenge questions and thereby stimulated a great deal of further work. These problems were solved in a later treatise, which in turn posed new mathematical problems to be solved; this was the beginning of a tradition of posing and solving problems that lasted for 150 years.

Seki Kōwa and Takebe Kenkō. One figure in seventeenth-century Japanese mathematics stands out far above all others, a genius who is frequently compared with Archimedes, Newton, and Gauss.⁶ His name was Seki Kōwa, and he was born around the year 1642, the year in which Isaac Newton was born in England. The stories told of him bear a great resemblance to similar stories told about other mathematical geniuses. For example, one of his biographers says that at the age of 5 Seki Kōwa pointed out errors in a computation that was being discussed by his elders. A very similar story is told about Gauss. Being the child of a samurai father and adopted by a noble family, Seki Kōwa had access to books. He was mostly self-educated in mathematics, having paid little attention to those who tried to instruct him; in this respect he resembles Newton. Like Newton, he served as an advisor on high finance to the government, becoming examiner of accounts to the lord of Kosshu. Unlike Newton, however, he was a popular teacher and physically vigorous. He became a shogunate samurai and master of ceremonies in the household of the

⁶ His biography suggests that the real comparison should be with Pythagoras, since he assembled a devoted following, and his followers were inclined to attribute results to him even when his direct influence could not be established. Newton and Gauss were not "people persons," and Gauss hated teaching.

Shogun. He died at the age of 66, leaving no direct heirs. His tomb in the Buddhist cemetery in Tokyo was rebuilt 80 years after his death by mathematicians of his school. His pedagogical activity earned him the title of *Sansei*, meaning *Arithmetical Sage*, a title that was carved on his tomb. Although he published very little during his lifetime, his work became known through his teaching activity, and he is said to have left copious notebooks.

Seki Kōwa made profound contributions to several areas of mathematics, in some cases anticipating results that were being obtained independently in Europe about this time. According to Mikami (1913, p. 160), he kept his technique a secret from the world at large; but apparently he confided it to his pupil Takebe Kenkō (Takebe Katahiro, 1664–1739). Some scholars say that Takebe Kenkō refused to divulge the secret, saying, “I fear that one whose knowledge is so limited as mine would tend to misrepresent its significance.” However, other scholars claim that Takebe Kenkō did write an exposition of the latter method, and that it amounts to the principles of cancellation and transposition. These two scholars, together with Takebe Kenkō’s brother, compiled a 20-volume encyclopedia, the *Taisei Sankyō* (*Great Mathematical Treatise*), containing all the mathematics known in their day.

Takebe Kenkō also wrote a book that is unique in its time and place, bearing the title *Tetsujutsu Sankyō* (roughly, *The Art of Doing Mathematics*, published in 1722), in which he speculated on the metaphysics of mathematical concepts and the kind of psychology needed to solve different types of mathematical problems (Murata, 1994, pp. 107–108).

In Japan, knowledge of the achievements of Western mathematicians became widespread in the late nineteenth century, while the flow of knowledge in the opposite direction has taken longer. A book entitled *The Theory of Determinants in the Historical Order of Development*, which is a catalog of papers on the subject with commentaries, was written by the South African mathematician Thomas Muir (1844–1934) in 1905. Although this book consists of four volumes totaling some 2000 pages, it does not mention Seki Kōwa, the true discoverer of determinants!

Other treatises. The book *Sampō Ketsugi-shō* (*Combination Book*, but it contains many results on areas and volumes that we now compute using calculus) was published in 1661 by Isomura Yoshinori (Isomura Kittoku), a student of a student of Mōri Shigeyoshi. Although Isomura is known to have died in 1710, his birth date is uncertain. The book was revised in 1684. Sawaguchi Kazuyuki, whose exact dates also are not known, wrote *Kokon Sampō-ki* (*Mathematics Ancient and Modern*) in 1671. This work is cited by Murata as the proof that *wasan* had developed beyond its Chinese origins.

The modern era in Japan. In the seventeenth century the Tokugawa shoguns had adopted a very strict policy vis-à-vis the West, one that could be enforced in an island kingdom such as Japan. Commercial contacts with the Dutch, however, resulted in some cultural penetration, and Western mathematical advances came to be known little by little in Japan. By the time Japan was opened to the West in the mid-nineteenth century, Japanese mathematicians were already aware of many European topics of investigation. In joining the community of nations for trade and politics, Japan also joined it intellectually. In the early nineteenth century, Japanese mathematicians were writing about such questions as the rectification of the ellipse, a subject of interest in Europe at the same period. By the end of the nineteenth century there were several Japanese mathematical journals publishing

(in European languages) mathematical work comparable to what was being done in Europe at the same period, and a few European scholars were already reading these journals to see what advances were being made by the Japanese. In the twentieth century the number of Japanese works being read in the West multiplied, and Japanese mathematicians such as Gorō Shimura (b. 1930), Shōshichi Kobayashi (b. 1932), and many others have been represented among the leaders in nearly every field of mathematics.

3. The Muslims

From the end of the eighth century through the period referred to as Medieval in European history, the Umayyad and Abbasid Caliphates, centered in what is now Spain and Iraq respectively, produced an artistically and scientifically advanced culture, with works on mathematics, physics, chemistry, and medicine written in Arabic, the common language of scholars throughout the Muslim world. Persian, Hebrew, and other languages were also used by scholars working in this predominantly Muslim culture. Hence the label “Islamic mathematics” that we prefer to use is only a rough description of the material we shall be discussing. It is convenient, like the label “Greek mathematics” used above to refer to works written in the culture where scholars mostly wrote in Greek.

3.1. Islamic science in general. The religion of Islam calls for prayers facing Mecca at specified times of the day. That alone would be sufficient motive for studying astronomy and geography. Since the Muslim calendar is lunar rather than lunisolar, religious feasts and fasts are easy to keep track of. Since Islam forbids representation of the human form in paintings, mosques are always decorated with abstract geometric patterns (see Özdural, 2000). The study of this *ornamental geometry* has interesting connections with the theory of transformation groups.

Hindu influences. According to Colebrooke (1817, pp. lxiv–lxv), in the year 773 CE, al-Mansur, the second caliph of the Abbasid dynasty, who ruled from 754 to 775, received at his court a Hindu scholar bearing a book on astronomy referred to in Arabic as *Sind-hind* (most likely, *Siddhanta*). Al-Mansur had this book translated into Arabic. No copies survive, but the book seems to have been the *Brahmasphutasiddhanta* mentioned above. This book was used for some decades, and an abridgement was made in the early ninth century, during the reign of al-Mamun (caliph from 813 to 833), by Muhammed ibn Musa al-Khwarizmi (ca. 780–850), who also wrote his own treatise on astronomy based on the Hindu work and the work of Ptolemy. Al-Mamun founded a “House of Wisdom” in Baghdad, the capital of his empire. This institution was much like the Library at Alexandria, a place of scholarship analogous to a modern research institute.

In the early days of this scientific culture, one of the main concerns of the scholars was to find and translate into Arabic as many scientific works as possible. The effort made by Islamic rulers, administrators, and merchants to acquire and translate Hindu and Hellenistic texts was prodigious. The works had first to be located, a job requiring much travel and expense. Next, they needed to be understood and adequately translated; that work required a great deal of labor and time, often involving many people. The world is much indebted to the scholars who undertook this work, for two reasons. First, some of the original works have been lost, and

only their Arabic translations survive.⁷ Second, the translators, inspired by the work they were translating, wrote original works of their own. The mechanism of this two-part process has been well described by Berggren (1990, p. 35):

Muslim scientists and patrons were the main actors in the acquisition of Hellenistic science inasmuch as it was they who initiated the process, who bore the costs, whose scholarly interests dictated the choice of material to be translated and on whom fell the burden of finding an intellectual home for the newly acquired material within the Islamic *dār al-‘ilm* (“abode of learning”).

We shall describe the two parts of the process as “acquisition” and “development.” The acquisitions were too many to be listed here. Some of the major ones were listed by Berggren (2002). They include Euclid’s *Elements*, *Data*, and *Phænomena*, Ptolemy’s *Syntaxis* (which became the *Almagest* as a result) and his *Geography*, many of Archimedes’ works and commentaries on them, and Apollonius’ *Conics*.

The development process as it affected the *Conics* of Apollonius was described by Berggren (1990, pp. 27–28). This work was used to analyze the astrolabe in the ninth century and to trisect the angle and construct a regular heptagon in the tenth century. It continued to be used down through the thirteenth century in the theory of optics, for solving cubic equations, and to study the rainbow. To the two categories that we have called acquisition and development Berggren adds the process of editing the texts to systematize them, and he emphasizes the very important role of mathematical philosophy or criticism engaged in by Muslim mathematicians. They speculated and debated Euclid’s parallel postulate, for example, thereby continuing a discussion that began among the ancient Greeks and continued for 2000 years until it was finally settled in the nineteenth century.

The scale of the Muslim scientific schools is amazing when looked at in comparison with the populations and the general level of economic development of the time. Here is an excerpt from a letter of the Persian mathematician al-Kashi (d. 1429) to his father, describing the life of Samarkand, in Uzbekistan, where the great astronomer Ulugh Beg (1374–1449), grandson of the conqueror Timur the Lame, had established his observatory (Bagheri, 1997, p. 243):

His Royal Majesty had donated a charitable gift... amounting to thirty thousand... dinars, of which ten thousand had been ordered to be given to students. [The names of the recipients] were written down; [thus] ten thousand-odd students steadily engaged in learning and teaching, and qualifying for a financial aid, were listed... Among them there are five hundred persons who have begun [to study] mathematics. His Royal Majesty the World-Conqueror, may God perpetuate his reign, has been engaged in this art... for the last twelve years.

⁷ Toomer (1984) points out that in the case of Ptolemy’s *Optics* the Arabic translation has also been lost, and only a Latin translation from the Arabic survives. As Toomer notes, some of the most interesting works were not available in Spain and Sicily, where medieval scholars went to translate Arabic and Hebrew manuscripts into Latin.

3.2. Some Muslim mathematicians and their works. Continuing with our list of the major writers and their works, we now survey some of the more important ones who lived and worked under the rule of the caliphs.

Muhammed ibn Musa al-Khwarizmi. This scholar translated a number of Greek works into Arabic but is best remembered for his *Hisab al-Jabr w'al-Mugabalah* (*Book of the Calculation of Restoration and Reduction*). The word *restoration* here (*al-jabr*) is the source of the modern word *algebra*. It refers to the operation of keeping an equation in balance by transferring a term from one side to the opposite side with the opposite sign. The word *reduction* refers to the cancellation of like terms or factors from the two sides of an equation. The author came to be called simply al-Khwarizmi, which may be the name of his home town (although this is not certain); this name gave us another important term in modern mathematics, *algorithm*.

The integration of intellectual interests with religious piety that we saw in the case of the Hindus is a trait also possessed by the Muslims. Al-Khwarizmi introduces his algebra book with a hymn of praise of Allah, then dedicates his book to al-Mamun:

That fondness for science, by which God has distinguished the Imam al-Mamun, the Commander of the Faithful . . . , that affability and condescension which he shows to the learned, that promptitude with which he protects and supports them in the elucidation of obscurities and in the removal of difficulties —has encouraged me to compose a short work on Calculating by (the rules of) Completion and Reduction, confining it to what is easiest and most useful in arithmetic, such as men constantly require in cases of inheritance, legacies, partition, law-suits, and trade, and in all their dealings with one another, or where the measuring of lands, the digging of canals, geometrical computation, and other objects of various sorts. . . My confidence rests with God, in this as in every thing, and in Him I put my trust. . . May His blessing descend upon all the prophets and heavenly messengers. [Rosen, 1831, pp. 3–4]

Thabit ibn-Qurra. The Sabian (star-worshipping) sect centered in the town of Haran in what is now Turkey produced an outstanding mathematician/astronomer in the person of Thabit ibn-Qurra (826–901). Being trilingual (besides his native Syriac, he spoke Arabic and Greek), he was invited to Baghdad to study mathematics. His mathematical and linguistic skills procured him work translating Greek treatises into Arabic, including Euclid's *Elements*. He was a pioneer in the application of arithmetic operations to ratios of geometric quantities, which is the essence of the idea of a real number. The same idea occurred to René Descartes (1596–1650) and was published in his famous work on analytic geometry. It is likely that Descartes drew some inspiration from the works of the fourteenth-century Bishop of Lisieux Nicole d'Oresme (1323–1382); Oresme, in turn, is likely to have read translations from the Arabic. Hence it is possible that our modern concept of a real number can be traced back to the genius of Thabit ibn-Qurra. He also wrote on mechanics, geometry, and number theory.

Abu-Kamil. Although nothing is known of the life of Abu-Kamil (ca. 850–93), he is the author of certain books on algebra, geometry, and number theory that had a marked influence on both Islamic mathematics and the recovery of mathematics in Europe. Many of his problems were reproduced in the work of the Leonardo of Pisa (Fibonacci, 1170–1226).

Abu'l-Wafa. Mohammad Abu'l-Wafa (940–998) was born in Khorasan (now in Iran) and died in Baghdad. He was an astronomer–mathematician who translated Greek works and commented on them. In addition he wrote a number of works on practical arithmetic and geometry. According to Rashid (1994), his book of practical arithmetic for scribes and merchants begins with the claim that it “comprises all that an experienced or novice, subordinate or chief in arithmetic needs to know” in relation to taxes, business transactions, civil administration, measurements, and “all other practices . . . which are useful to them in their daily life.”

Al-Biruni. Abu Arrayhan al-Biruni (973–1048), was an astronomer, geographer, and mathematician who as a young man worked out the mathematics of maps of Earth. Civil wars in the area where he lived (Uzbekistan and Afghanistan) made him into a wanderer, and he came into contact with astronomers in Persia and Iraq. He was a prolific writer. According to the *Dictionary of Scientific Biography*, he wrote what would now be well over 10,000 pages of texts during his lifetime, on geography, geometry, arithmetic, and astronomy.

Omar Khayyam. The Persian mathematician Omar Khayyam was born in 1044 and died in 1123. He is thought to be the same person who wrote the famous skeptical and hedonistic poem known as the *Rubaiyat* (*Quatrains*), but not all scholars agree that the two are the same. Since he lived in the turbulent time of the invasion of the Seljuk Turks, his life was not easy, and he could not devote himself wholeheartedly to scholarship. Even so, he advanced algebra beyond the elementary linear and quadratic equations that one can find in al-Khwarizmi's book and speculated on the foundations of geometry. He explained his motivation for doing mathematics in the preface to his *Algebra*. Like the Japanese *wasanists*, he was inspired by questions left open by his predecessors. Also, as with al-Khwarizmi, this intellectual curiosity is linked with piety and thanks to the patron who supported his work.

In the name of God, gracious and merciful! Praise be to God, lord of all Worlds, a happy end to those who are pious, and ill-will to none but the merciless. May blessings repose upon the prophets, especially upon Mohammed and all his holy descendants.

One of the branches of knowledge needed in that division of philosophy known as mathematics is the science of completion and reduction, which aims at the determination of numerical and geometrical unknowns. Parts of this science deal with certain very difficult introductory theorems, the solution of which has eluded most of those who have attempted it. . . I have always been very anxious to investigate all types of theorems and to distinguish those that can be solved in each species, giving proofs for my distinctions, because I know how urgently this is needed in the solution of difficult problems. However, I have not been able to find time to complete this work, or to concentrate my thoughts on it, hindered as I have been by troublesome obstacles. [Kasir, 1931, pp. 43–44]

Al-Tusi. Nasir al-Din al-Tusi (1201–1274) had the misfortune to live during the time of the westward expansion of the Mongols, who subdued Russia during the 1240s, then went on to conquer Baghdad in 1258. Al-Tusi himself joined the Mongols and was able to continue his scholarly work under the new ruler Hulegu, grandson of Genghis Khan. Hulegu, who died in 1265, conquered and ruled Iraq and Persia over the last decade of his life, taking the title *Ilkhan* when he declared himself ruler of Persia. A generation later the Ilkhan rulers converted from Buddhism to Islam. Hulegu built al-Tusi an observatory at Maragheh, a city in the Azerbaijan region of Persia that Hulegu had made his seat of government. Here al-Tusi was able to improve on the earlier astronomical theory of Ptolemy, in connection with which he developed both plane and spherical trigonometry into much more sophisticated subjects than they had been previously. Because of his influence, the loss of Baghdad was less of a blow to Islamic science than it would otherwise have been. Nevertheless, the constant invasions had the effect of greatly reducing the vitality and the quantity of research. Al-Tusi played an important role in the flow of mathematical ideas back into India after the Muslim invasion of that country; it was his revised and commented edition of Euclid's *Elements* that was mainly studied (de Young, 1995, p. 144).

4. Europe

As the western part of the world of Islam was growing politically and militarily weaker because of invasion and conquest, Europe was entering on a period of increasing power and vigor. One expression of that new vigor, the stream of European mathematical creativity that began as a small rivulet 1000 years ago, has been steadily increasing until now it is an enormous river and shows no sign of subsiding.

4.1. Monasteries, schools, and universities. From the sixth to the ninth centuries a considerable amount of classical learning was preserved in the monasteries in Ireland, which had been spared some of the tumult that accompanied the decline of Roman power in the rest of Europe. From this source came a few scholars to the court of Charlemagne to teach Greek and the quadrivium (arithmetic, geometry, music, and astronomy) during the early ninth century. Charlemagne's attempt to promote the liberal arts, however, encountered great obstacles, as his empire was divided among his three sons after his death. In addition, the ninth and tenth centuries saw the last waves of invaders from the north—the Vikings, who disrupted commerce and civilization both on the continent and in Britain and Ireland until they became Christians and adopted a settled way of life. Despite these obstacles, Charlemagne's directive to create cathedral and monastery schools had a permanent effect, contributing to the synthesis of observation and logic known as modern science.

Gerbert. In the chaos that accompanied the breakup of the Carolingian Empire and the Viking invasions the main source of stability was the Church. A career in public life for one not of noble birth was necessarily an ecclesiastical career, and church officials had to play both pastoral and diplomatic roles. That some of them also found time for scholarly activity is evidence of remarkable talent. Such a talent was Gerbert of Aurillac (ca. 940–1002). He was born to lower-class but free parents in south-central France. He benefited from Charlemagne's decree that monasteries and cathedrals must have schools and was educated in Latin grammar

at the monastery of St. Gerald in Aurillac. Throughout a vigorous career in the Church that led to his coronation as Pope Sylvester II in the year 999 he worked for a revival of learning, both literary and scientific. (He was not a successful clergyman or pope. He got involved in the politics of his day, offended the Emperor, and was suspended from his duties as Archbishop of Reims by Pope Gregory V in 998. He was installed as pope by the 18-year-old Emperor Otto II in 999, but after only three years both he and Otto were driven from Rome by a rebellion. Otto died trying to reclaim Rome, and Sylvester II died shortly afterward.)

4.2. The high Middle Ages. By the midtwelfth century European civilization had absorbed much of the learning of the Islamic world and was nearly ready to embark on its own explorations. This was the zenith of papal power in Europe, exemplified by the ascendancy of the popes Gregory VII (1073–1085) and Innocent III (1198–1216) over the emperors and kings of the time. The Emperor Frederick I, known as Frederick Barbarossa because of his red beard, who ruled the empire from 1152 to 1190, tried to maintain the principle that his power was not dependent on the Pope, but was ultimately unsuccessful. His grandson Frederick II (1194–1250) was a cultured man who encouraged the arts and sciences. To his court in Sicily he invited distinguished scholars of many different religions, and he corresponded with many others. He himself wrote a treatise on the principles of falconry. He was in conflict with the Pope for much of his life and even tried to establish a new religion, based on the premise that “no man should believe aught but what may be proved by the power and reason of nature,” as the papal document excommunicating him stated.

4.3. Authors and works. A short list of European mathematicians prominent in their time from the twelfth through sixteenth centuries begins in the empire of Frederick II.

Leonardo of Pisa. Leonardo says in the introduction to his major book that he accompanied his father on an extended commercial mission in Algeria with a group of Pisan merchants. There, he says, his father had him instructed in the Hindu–Arabic numerals and computation, which he enjoyed so much that he continued his studies while on business trips to Egypt, Syria, Greece, Sicily, and Provence. Upon his return to Pisa he wrote a treatise to introduce this new learning to Italy. The treatise, whose author is given as “Leonardus filius Bonaccij Pisani,” that is, “Leonardo, son of Bonaccio of Pisa,” bears the date 1202. In the nineteenth century Leonardo’s works were edited by the Italian nobleman Baldassare Boncompagni (1821–1894), who also compiled a catalog of locations of the manuscripts (Boncompagni, 1854). The name Fibonacci by which the author is now known seems to have become generally used only in the nineteenth century.

Jordanus Nemorarius. The works of Archimedes were translated into Latin in the thirteenth century, and his work on the principles of mechanics was extended. One of the authors involved in this work was Jordanus Nemorarius. Little is known about this author except certain books that he wrote on mathematics and statics for which manuscripts still exist dating to the actual time of composition.

Nicole d’Oresme. One of the most distinguished of the medieval philosophers was Nicole d’Oresme, whose clerical career brought him to the office of Bishop of Lisieux in 1377. D’Oresme had a wide-ranging intellect that covered economics, physics, and mathematics as well as theology and philosophy. He considered the motion of

physical bodies from various points of view, formulated the Merton rule of uniformly accelerated motion (named for Merton College, Oxford), and for the first time in history explicitly used one line to represent time, a line perpendicular to it to represent velocity, and the area under the graph (as we would call it) to represent distance.

Regiomontanus. The work of translating the Greek and Arabic mathematical works went on for several centuries. One of the last to work on this project was Johann Müller of Königsberg (1436–1476), better known by his Latin name of Regiomontanus, a translation of Königsberg (King's Mountain). Although he died young, Regiomontanus made valuable contributions to astronomy, mathematics, and the construction of scientific measuring instruments. In all this he bears a strong resemblance to al-Tusi, mentioned above. He studied in Leipzig while a teenager, then spent a decade in Vienna and the decade following in Italy and Hungary. The last five years of his life were spent in Nürnberg. He is said to have died of an epidemic while in Rome as a consultant to the Pope on the reform of the calendar.

Regiomontanus checked the data in copies of Ptolemy's *Almagest* and made new observations with his own instruments. He laid down a challenge to astronomy, remarking that further improvement in theoretical astronomy, especially the theory of planetary motion, would require more accurate measuring instruments. He established his own printing press in Nürnberg so that he could publish his works. These works included several treatises on pure mathematics. He established trigonometry as an independent branch of mathematics rather than a tool in astronomy. The main results we now know as plane and spherical trigonometry are in his book *De triangulis omnimodis*, although not exactly in the language we now use.

Chuquet. The French Bibliothèque Nationale is in possession of the original manuscript of a comprehensive mathematical treatise written at Lyons in 1484 by one Nicolas Chuquet. Little is known about the author, except that he describes himself as a Parisian and a man possessing the degree of Bachelor of Medicine. The treatise consists of four parts: a treatise on arithmetic and algebra called *Triparty en la science des nombres*, a book of problems to illustrate and accompany the principles of the *Triparty*, a book on geometrical mensuration, and a book of commercial arithmetic. The last two are applications of the principles in the first book.

Luca Pacioli. Written at almost the same time as Chuquet's *Triparty* was a work called the *Summa de arithmetica, geometrica, proportioni et proportionalitate* by Luca Pacioli (or Paciuolo) (1445–1517). Since Chuquet's work was not printed until the nineteenth century, Pacioli's work is believed to be the first Western printed work on algebra. In comparison with the *Triparty*, however, the *Summa* seems less original. Pacioli has only a few abbreviations, such as *co* for *cosa*, meaning *thing* (the unknown), *ce* for *censo* (the square of the unknown), and *æ* for *æquatur* (equals). Despite its inferiority to the *Triparty*, the *Summa* was much the more influential of the two books, because it was published. It is referred to by the Italian algebraists of the early sixteenth century as a basic source.

Leon Battista Alberti. In art the fifteenth century was a period of innovation. In an effort to give the illusion of depth in two-dimensional representations some artists looked at geometry from a new point of view, studying the projection of two- and three-dimensional shapes in two dimensions to see what properties were preserved

and how others were changed. A description of such a procedure, based partly on the work of his predecessors, was given by Leon Battista Alberti (1404–1472) in a treatise entitled *Della pittura*, published posthumously in 1511.

Sixteenth-century Italy produced a group of sometimes quarrelsome but always brilliant algebraists, who worked to advance their science for the sheer pleasure of making new mathematical achievements. As happened in Japan a century later, each new advance brought a challenge for further progress.

Scipione del Ferro. A method of solving a cubic equations was discovered by a lector (reader, that is, a tutor) at the University of Bologna, Scipione del Ferro (1465–1525), around the year 1500. He communicated this discovery to another mathematician, Antonio Maria Fior (dates unknown), who then used the knowledge to win mathematical contests.

Niccolò Tartaglia. Fior met his match in 1535, when he challenged Niccolò Fontana (1500–1557) of Brescia, known as Tartaglia (the Stammerer) because a wound he received as a child when the French overran Brescia in 1512 left him with a speech impediment. Tartaglia had also discovered how to solve certain cubic equations and so won the contest.

Girolamo Cardano. A brilliant mathematician and gambler, who became rector of the University of Padua at the age of 25, Girolamo Cardano (1501–1576) was writing a book on mathematics in 1535 when he heard of Tartaglia's victory over Fior. He wrote to Tartaglia asking permission to include this technique in his work. Tartaglia at first refused, hoping to work out all the details of all cases of the cubic and write a treatise himself. According to his own account, Tartaglia confided the secret of one kind of cubic to Cardano in 1539, after Cardano swore a solemn oath not to publish it without permission and gave Tartaglia a letter of introduction to the Marchese of Vigevano. Tartaglia revealed a rhyme by which he had memorized the procedure.

Tartaglia did not claim to have given Cardano any proof that his procedure works. It was left to Cardano himself to find the demonstration. Cardano kept his promise not to publish this result until 1545. However, as Tartaglia delayed his own publication, and in the meantime Cardano had discovered the solution of other cases of the cubic himself and had also heard that del Ferro had priority anyway, he published the result in his *Ars magna* (*The Great Art*), giving credit to Tartaglia. Tartaglia was furious and started a bitter controversy over Cardano's alleged breach of faith.

Ludovico Ferrari. Cardano's student Ludovico Ferrari (1522–1565) worked with him in the solution of the cubic, and between them they had soon found a way of solving certain fourth-degree equations.

Rafael Bombelli. In addition to the mathematicians proper, we must also mention an engineer in the service of an Italian nobleman. Rafael Bombelli (1526–1572) is the author of a treatise on algebra that appeared in 1572. In the introduction to this treatise we find the first mention of Diophantus in the modern era. Bombelli said that, although all authorities are agreed that the Arabs invented algebra, he, having been shown the work of Diophantus, credits the invention to the latter. In making sense of what his predecessors did he was one of the first to consider the

square root of a negative number and to formulate rules for operating with such numbers.

The work being done in Italy did not escape the notice of French and British scholars of the time, and important mathematical works were soon being produced in those two countries.

François Viète. A lawyer named François Viète (1540–1603), who worked as tutor in a wealthy family and later became an advisor to Henri de Navarre (who became the first Bourbon king, Henri IV, in 1598), found time to study Diophantus and to introduce his own ideas into algebra. His book *Artis analyticae praxis* (*The Practice of the Analytic Art*) contained some of the notational innovations that make modern algebra much less difficult than the algebra of the sixteenth century.

Girard Desargues. Alberti's ideas on projection were extended by the French architect and engineer Girard Desargues (1591–1661), who studied the projections of figures in general and the conic sections in particular.

John Napier. In the late sixteenth century the problem of simplifying laborious multiplications, divisions, root extractions, and the like, was attacked by the Scottish laird John Napier, Baron of Murchiston (1550–1617). His work consisted of two parts, a theoretical part, based on a continuous geometric model, and a computational part, involving a discrete (tabular) approximation of the continuous model. The computational part was published in 1614. However, Napier hesitated to publish his explanation of the theoretical foundation. Only in 1619, two years after his death, did his son publish an English translation of Napier's theoretical work under the title *Mirifici logarithmorum canonis descriptio* (*A Description of the Marvelous Law of Logarithms*). This subject, although aimed at a practical end, turned out to have enormous value in theoretical studies as well.

The European colonies. Wherever Europeans went during their great age of expansion, science and mathematics followed once the new lands were settled and acquired political stability and a certain level of economic prosperity. Like the mathematics of Europe proper, the story of this "colonial" mathematics is too large to fit into the present volume, and so we shall, with regret, omit South America and South Africa from the story and concentrate on the origins of mathematics in Mexico, the United States, Canada, Australia, and New Zealand.

5. North America

During the American colonial period and for nearly a century after the founding of the United States, mathematical research in North America was extremely limited. Educational institutions were in most cases directed toward history, literature, and classics, the major exception being the academy at West Point, which became the United States Military Academy in 1802. Modeling itself consciously on the École Polytechnique, the Academy taught engineering and applied mathematics.⁸ For most of the period up to 1875 there were no professional journals devoted entirely to mathematics and no mathematical societies of any size. A period of rapid growth began in the 1870s, coinciding with the closing of the American frontier. By 1900 a respectable school of American mathematical researchers existed, although it was

⁸ Rensselaer Polytechnic Institute was founded to teach engineering in 1824, and civil engineering was taught at the University of Vermont as early as 1829.

still puny compared with the schools in Germany, Britain, France, and Italy. Even as late as 1940, only about half a dozen mathematical journals were published in the United States. The United States vaulted to a position of world leadership in mathematics following World War II, and it has remained among the strongest nations in this area, thanks to its possession of a powerful university system and equally well-developed professional organizations such as the American Mathematical Society, the Mathematical Association of America, the Society for Industrial and Applied Mathematics, and the National Council of Teachers of Mathematics, together with over 100 professional journals devoted to mathematics in general or specific areas within it.

5.1. The United States and Canada before 1867. Until the late nineteenth century most of the mathematics done in North America was purely practical, and to find more than one or two examples of its practitioners we shall have to leave mathematics proper and delve into related areas. Nevertheless, one can find a few examples of Americans who practiced mathematics for its own sake, even in the eighteenth century.

David Rittenhouse. Like his younger brother Benjamin (1740–1825), David Rittenhouse (1732–1796) was primarily a manufacturer of compasses and clocks. He made two compasses for George Washington. He also got involved in surveying and in 1763 helped to settle a border dispute between William Penn and Lord Baltimore. He became the first director of the United States Mint by appointment of President Washington in 1792, and he became president of the American Philosophical Society in 1791, after the death of Benjamin Franklin. According to Homann (1987), he was self-taught in mathematics, but enjoyed calculation very much and so was able to read Newton's *Principia* on his own. He developed a continued-fraction method of approximating the logarithm of a positive number, described in detail by Homann. Like the Japanese tradition of challenge problems, some of Rittenhouse's papers asked for proofs of results the author himself had not been able to supply. In one case this challenge was taken up by Nathaniel Bowditch (discussed below).

Robert Adrain. An immigrant of great mathematical talent—he came to the United States from his native Ireland after being wounded by friendly fire in the rebellion of 1798—was Robert Adrain (1775–1843). He taught at Princeton until 1800, when he moved to York, Pennsylvania; in 1804 he moved again, to Reading, Pennsylvania. He contributed to, and in 1807 became editor of, the *Mathematical Correspondent*, the first mathematical research journal in the United States. Parshall (2000, p. 381) has noted that even as late as 1874 “[t]here were no journals in the United States devoted to mathematical research, and, in fact, up to that time all attempts to sustain such publication outlets had failed almost immediately.” The *Mathematical Correspondent* appears to have ended with the first issue of Vol. 2, that is, the first one edited by Adrain. In an interesting article on the original editor of the *Mathematical Correspondent*, George Baron (b. 1769, date of death unknown), V. Fred Rickey notes that perhaps it may not have been merely the American ignorance of mathematics that led to an early demise for this journal. Rickey points out that the journal had 347 subscribers and published 487 copies of its first issue, but that an article in *The Analyst* in 1875 (2, No. 5, 131–138) by one David S. Hart contains the following interesting comment:

The writer has a copy of No. 2. stitched in a blue cover, on which is an advertisement of a Lecture delivered in New York by G. Baron, which contains (as he says) "a complete refutation of the false and spurious principles, ignorantly imposed on the public, in the 'New American Practical Navigator,' written by N. Bowditch and published by E.M. Blunt." The sub-editors endorsing the above say, "We agree with the author that he has shown in the most incontrovertible manner, that the principles on which the 'New American Practical Navigator' is founded, are universally false, and gross impositions on the public."

Since Bowditch was, next to Adrain, the strongest mathematician in the country at the time, this sort of internecine feuding could only have been harmful to the development of a community of mathematicians. Rickey's article can be found by following links from the following website.

<http://www.dean.usma.edu/math/people/rickey/>

Adrain is best remembered for discovering, independently of Legendre and Gauss, the theory of least-squares and the normal (Gaussian) distribution. However, given the low state of science in general in the United States, it is not surprising that no one in Europe noticed Adrain's work. Kowalewski (1950, pp. 84–85) notes that the Göttingen astronomer Tobias Mayer (1723–1762) had used a similar method as early as 1748.

Commerce requires a certain amount of mathematics and astronomy to meet the needs of navigation, and all the early American universities taught dialing (theory of the sundial), astronomy, and navigation. These subjects were standard, long-known mathematics, a great contrast to the rapid pace of innovation in Europe at this period. Nevertheless, to write the textbooks of navigation and calculate the tides a year in advance required some ability. It is remarkable that this knowledge was acquired by two Americans who were not given even the limited formal education that could be obtained at an American university. Although neither was a mathematician in the strict sense, both of them understood and used the mathematics of astronomy.

Benjamin Banneker. In the fall of 1791 the Baltimore publishing house of William Goddard and James Angell published a book bearing the title *Banneker's Almanac and Ephemeris for the Year of our Lord 1792*. . . . The author, Benjamin Banneker (1731–1806), was 60 years old at the time, the only child of parents of African descent⁹ who had left him a small parcel of land as an inheritance. For most of his life Banneker lived near Baltimore, struggling as a poor farmer with a rudimentary formal education. Nevertheless, he acquired a reputation for cleverness due to his skill in arithmetic. In middle age he made the acquaintance of the Ellicotts, a prominent local family, who lent him a few books on astronomy. From these meager materials Banneker was able to construct an almanac for the year 1791. Encouraged by this success, he prepared a similar almanac for 1792. In that year the Ellicotts put him in contact with James McHenry (who had been Surgeon General of the American Army during the Revolutionary War). McHenry wrote to the editors:

⁹ Banneker's grandmother was an Englishwoman who married one of her slaves. Their daughter Mary, Banneker's mother, also married a slave, who had the foresight to purchase a farm jointly in his own name and in the name of his son Benjamin.

[H]e began and finished [this almanac] without the least information or assistance from any person, or other books than those I have mentioned; so that whatever merit is attached to his present performance is exclusively and peculiarly his own.

Banneker's *Almanac* was published and sold all over the United States in the decade from 1792 until 1802. The contents of the *Almanac* are comparable with those of other almanacs that have been published in the United States: On alternate pages one finds calendars for each week or month, giving the phases of the Moon, the locations of the planets and bright stars visible during the period in question, and the times of sunrise, high and low tides, and conjunctions and oppositions of planets. Recognition came late to Banneker. The money he earned from his *Almanac* gave him some leisure in his old age, and his name was praised by Pitt in Parliament and by Condorcet before the French Academy of Sciences.

African-American mathematicians. Although the antislavery movement had begun in Banneker's time, African Americans were to endure two more generations of slavery followed by three generations of institutionalized, legalized discrimination and disenfranchisement before the civil rights movement gained sufficient strength to open to them the opportunities that a white American of very modest means could expect. It is therefore no wonder that very few African Americans became noted scholars. Nevertheless, the scientific creativity of African Americans has been a significant factor in the economic life of the United States, as can be seen, for example, in the book of James (1989). The first African American to obtain a doctorate in mathematics was Elbert Cox (1895–1969), who became a professor at Howard University after obtaining the doctorate at Cornell in 1925, one of only 28 doctorates awarded to Americans (of any color) that year. The first African-American women to receive the doctorate in mathematics, both of them in 1949, were Marjorie Lee Brown (1914–1979) and Evelyn Boyd Granville (b. 1924). Brown was a differential topologist who received her degree at the University of Michigan and taught at North Carolina Central University. Granville received the Ph.D. from Yale University and worked in the space program during the 1960s. She later taught at California State University in Los Angeles.

The number of African Americans choosing to enter mathematics and science is still comparatively small. In fact, the author of an article entitled "Black Women Ph.D.'s in Mathematics" in the 1980s was able to interview *all* of the people described in the title who were still alive. A career in research, after all, requires a long apprenticeship, during which financial support must be provided either by family, by extra work, or by grants and loans. For people who do not come from wealthy families, other careers, promising earlier financial rewards, are likely to seem more attractive. Undoubtedly, if the average income of African Americans were higher, more of them would choose scientific careers. Lest these comments seem unduly pessimistic, it should be noted that a conference devoted to the research of African Americans in 1996 brought together 79 African-American mathematicians (Dean, 1996).

Nathaniel Bowditch. Benjamin Banneker was about 40 years old and living in obscurity near Baltimore when Nathaniel Bowditch (1773–1838) was born in Salem, Massachusetts. His ancestors had been shipbuilders but had accumulated no substantial amount of money by this trade. His father abandoned it and became a

cooper, a trade that barely provided for his family of seven children. Nathaniel received only a rudimentary public education before being apprenticed to a ship chandler at the age of 10. Twelve years later, when Banneker's *Almanac* had been published for only a year or two, he signed on board a ship and, like Banneker, used his few intervals of leisure to study mathematics and astronomy. Bowditch was a natural teacher who enthusiastically shared his knowledge of navigation with his shipmates. With his aptitude for mathematics, he managed to get through Newton's *Principia*, learning a considerable amount of Latin on the way. Later he taught himself French, which had displaced Latin as the language of science as a result of the pre-eminence of French mathematicians and scientists.

Bowditch first gained a scholarly reputation by pointing out errors in the standard navigational tables. His abilities immediately attracted interest, and his *Practical Navigator*, first published in 1800, gained him wide recognition¹⁰ while he was still in his twenties. Bowditch became a member of the American Academy of Arts and Letters, and in 1818 was elected a member of the Royal Society. With recognition came leisure time to devote to purely scholarly pursuits, a luxury denied to Banneker in his most vigorous years. For the last quarter-century of his life Bowditch labored on his monumental translation and commentary of the *Mécanique céleste* by Pierre-Simon Laplace (1749-1827). This work amounts really to a complete rewriting of Laplace's treatise, which shows the effects of a pronounced stinginess with ink and paper. Bowditch filled in all the missing details of arguments that Laplace had merely waved his hand at, not having the patience to write down arguments that had sometimes taken him weeks to discover. These pursuits brought Bowditch international fame, and he died covered with honors. The *American Journal of Science* published his obituary with a portrait of him in a classical Roman tunic which it is unlikely he ever actually wore.

5.2. The Canadian Federation and post Civil War United States. The end of the American Civil War in 1865 was followed closely by the founding of the Canadian Federation in 1867. The Federation was the result of the North America Act, which reserved some constitutional controls for Britain. Full independence came in 1982. From that time on, both countries experienced a cultural flowering, which included advances in mathematics. Americans and Canadians began to go to Europe to learn advanced mathematics. This early generation of European-trained mathematicians generally found no incentive to continue research upon returning home. However, they at least made the curriculum more sophisticated and prepared the way for the next generation.

In Europe there were more Ph.D. mathematicians being produced than the universities could absorb. Most of these entered other professions, but a few emigrated across the Atlantic. A scholarly coup was scored by Johns Hopkins University, which opened in 1876 with a first-rate mathematician on board, James Joseph Sylvester. Despite being 62 years old, Sylvester was still a creative algebraist, whose presence in America attracted international attention. One of his first acts was to found the first mathematical research journal in the United States, the *American Journal of Mathematics*. The founding of this journal had been suggested by William Edward Story (1850-1930), one of many Americans who went abroad to get the Ph.D. degree but, atypically, continued to do mathematical research after returning to the United States. Before Johns Hopkins was founded, there had been a few graduate

¹⁰ And apparently some detractors associated with the *Mathematical Correspondent* (see above).

programs in mathematics in places such as Harvard and the University of Michigan, but now such programs began to multiply. Bryn Mawr College opened in the mid-1880s with a graduate program in mathematics. The founding of Clark University in Worcester, Massachusetts and the University of Chicago in the late 1880s and early 1890s promised that the United States would soon begin to make respectable contributions to mathematical research. An account of this development giving the details of the mathematical areas studied in American universities can be found in the article by David Rowe (1997). A review of a number of professional “self-studies” made by American mathematicians can be found in the article by Karen Hunger Parshall (2000); both of these articles contain extensive bibliographies on the development of mathematics in the United States. We now continue our list of prominent mathematicians.

George William Hill. The mathematical side of astronomy, known as celestial mechanics, was pursued in the United States by the Canadian Simon Newcomb, who is discussed below, and by George William Hill (1838–1914). Hill worked for a time at the Nautical Almanac Office in Cambridge, Massachusetts, but was perfectly content to work in isolation at his home in Nyack, New York, most of his life. His work on the motion of the Moon was so profound that it received extravagant praise from Henri Poincaré (1854–1912), one of the greatest mathematicians of the late nineteenth and early twentieth centuries. In a paper on the motion of the lunar perigee published in the Swedish journal *Acta mathematica* in 1886, Hill derived a differential equation that bears his name and even today continues to generate new work.

Although the rise of the United States to a position of world leadership in mathematics after World War II was partly the result of the turbulence of the 1930s and 1940s, which drove many of the best European intellectuals to seek refuge far from the dangers that threatened them in their homelands, one should not think that the country was intellectually backward before that time. Americans had made significant contributions to algebra and logic in the nineteenth century, and in the early twentieth century a number of Americans achieved worldwide fame for their mathematical contributions. We mention only two here.

George David Birkhoff. Harvard professor George David Birkhoff (1884–1944) made contributions to differential equations, difference equations, ergodic theory, and mathematical physics (the kinetic theory of gases, in which the ergodic theorem plays a role, quantum mechanics, and relativity). He was held in such high esteem that a crater on the Moon now bears his name.

Norbert Wiener. An early prodigy who graduated from high school at age 11 and received the doctoral degree at age 18, despite having changed universities and majors more than once, Norbert Wiener (1894–1964) contributed to harmonic analysis, probability, quantum mechanics, and cybernetics, of which he was one of the founders. (The name comes from the Greek word *kybernetēs*, meaning a ship’s captain or pilot.)

Like American schools of the same period, English-language Canadian institutions of higher learning tended to rely on British textbooks such as those of Charles Hutton (1737–1823, a professor at the Military School in Woolwich). In French Canada there was a long tradition of educational institutions, and a French

calculus text written by Abbé Jean Langevin, who was to become Bishop of Rimouski in 1867, was published in 1848. For Canadians, as for Americans, the importance of research as an activity of the mathematics professor arose only after the founding of Johns Hopkins University in 1876. In fact, the early volumes of the *American Journal of Mathematics* contain articles by two Canadians, J. G. Glashan (1844–1932), superintendent of schools in Ottawa, and G. Paxton Young (1818–1889), a professor of philosophy at the University of Toronto.

Simon Newcomb. An outstanding nineteenth-century Canadian mathematician was Simon Newcomb (1835–1909), a native of Nova Scotia who taught school in a number of places in the United States before procuring a job at the Nautical Almanac Office in Cambridge, Massachusetts, where he attended Harvard. He eventually became director of the Naval Observatory in Washington, and after 1884 professor of mathematics at Johns Hopkins.

H. S. M. Coxeter. The geometer Harold Scott MacDonald Coxeter (1907–2003), a native of Britain, emigrated to Canada in 1936 and played a leading role in Canadian research in symmetry groups and symmetric geometric objects of all kinds. His work on tessellations inspired many famous paintings by the Dutch artist Maurits Escher (1898–1972).

John Synge. Although, strictly speaking, he counts as an Irish mathematician, who was born in Dublin and died there, John Synge (1897–1995) taught at the University of Toronto from 1920 to 1925 and again during the 1930s. From 1939 until 1948 he worked in the United States before returning to Ireland. He is listed here because of his daughter, Cathleen Synge Morawetz, who is discussed below.

John Charles Fields. One of the best-remembered Canadian mathematicians, John Charles Fields (1863–1932), was a native of Hamilton, Ontario. He received the Ph.D. from Johns Hopkins in 1887 and studied in Europe during the 1890s. In 1902 he became a professor at the University of Toronto. He wrote one book (on algebraic functions). Like many other mathematicians on the intellectual periphery of Europe, much of his activity was devoted to encouraging research in his native country. In the last few years of his life he established the Fields Medals, the highest international recognition for mathematicians, which are awarded at the quadrennial International Congress of Mathematicians. Beginning in 1936, when two awards were given, then resuming in 1950, the Fields Medals have by tradition been awarded to researchers early in their careers. As of 2002 about 40 mathematicians had been so honored, among them natives of China, Japan, New Zealand, the former Soviet Union, many European countries, and the United States.

Cecilia Krieger Dunaj. Canada has always taken in those fleeing oppression elsewhere, and some of these refugees have become prominent mathematicians. One example is Cypra Cecilia Krieger Dunaj (1894–1974), who studied mathematical physics at the University of Vienna before coming to Toronto in 1920, where she entered the university and took courses given by John Synge and John Fields. In 1930 she became the first woman to receive the doctoral degree in mathematics at a Canadian university (Toronto) and only the third woman to receive a doctoral degree in Canada.

Abraham Robinson. Among the mathematicians that the turbulent twentieth century condemned to wander the world was Abraham Robinson (1918–1974). He was born in what was then Germany and is now Poland, but emigrated with his family to Jerusalem in 1933, when the Nazis came to power in Germany. In 1940 he was studying in Paris, but evacuated to London when Paris fell to the Nazi invasion. After obtaining the Ph.D. and teaching in Britain for a few years, he spent six of his most productive years at the University of Toronto, beginning in 1951. While there he produced several Ph.D. students in mathematical logic. He left Toronto in 1957 to return to Jerusalem, but eventually moved to California and finally to Yale. The story of his Toronto years can be found in the article by Dauben (1996).

Cathleen Morawetz. Cathleen Synge Morawetz (b. 1923 in Toronto) is the daughter of John Synge. She attended the University of Toronto during World War II and then obtained the master's degree at the Massachusetts Institute of Technology in 1946. For a dissertation in mathematical physics she received the doctoral degree at New York University in 1951. Her subsequent career was very distinguished. She became associate director of the Courant Institute of Mathematical Sciences in 1978. In 1995–1996 she was president of the American Mathematical Society, the second woman to hold this post. In 1998 she was awarded the National Medal of Science, the highest scientific honor bestowed by the United States.

5.3. Mexico. The area that is now Mexico was the first part of the North American mainland to be colonized by Europeans and was the site of the first university in North America, the Universidad Real y Pontificia de México, founded in 1551. Unfortunately, the history of mathematics in modern Mexico has not been thoroughly studied, despite the fact that the mathematics of the earlier rulers of this part of the world, the Aztecs and Maya, has received quite a bit of scholarly attention. The present discussion amounts to a summary of the article by A. Garciadiego (2002), in which the author remarks that “the professionalization of the history of mathematics in Mexico is comparatively recent.”

The Royal University opened its doors just two years after its founding, offering a curriculum that was essentially medieval, consisting of theology, law, and related subjects. The first technical and scientific studies came more than a century later, with the establishment of a chair of astrology and mathematics. Unfortunately, some important scientific works were on the Index Librorum Prohibitorum, and Catholics were forbidden to read them.¹¹ Among these books were the works of Galileo and Newton, so that no real progress in science was to be expected.

Mexico became an independent country in 1821. An attempt by the French Emperor Louis Napoleon to make Mexico part of a renewed French Empire by establishing the puppet emperor Maximilian in 1864 soon failed. The French army left, and Maximilian was executed in 1867. The new leaders of Mexico sought to establish intellectual freedom that would incorporate material progress based on science. Despite these intentions, the University was closed at various times for political reasons. The National University of Mexico opened in 1910 and was accompanied by preparatory schools and schools for advanced studies. The most prominent name in the advance of mathematics and science in Mexico was Sotero Prieto (1884–1935), who taught advanced mathematics and physics and advocated

¹¹ It is interesting that astrology, belief in which is listed as a sin in the Catholic catechism, was not only permitted, but actually encouraged. The Index was not officially abolished until 1966, long after it had ceased to be taken seriously by either the faithful or the clergy.

the use of history in teaching. He is quoted as saying, "The history of a science clarifies the origins of its fundamental concepts and exhibits the evolution of its methods" (Garciadiago, 2002, p. 259).

Four years after the death of Prieto the Department of Mathematics was established as part of the Faculty of Sciences of the University, now known as the Universidad Nacional Autónoma de México. At this point, it could be said that the University had reached academic maturity. Seminars on current research opened, including one on scientific and philosophical problems. Foreign scholars came there to visit, and graduates from the University were able to find admission to first-rate universities in other countries. Upon his retirement from Princeton University in 1954, the distinguished topologist Solomon Lefschetz (1884–1972) accepted a position at the University of Mexico and began sending students to the graduate program at Princeton.¹² An amusing anecdote revealing the relations between the uninhibited Lefschetz and the Mexicans was reported by his student Gian-Carlo Rota (1932–1999). (See Rota's article 1989.)

6. Australia and New Zealand

Because of their proximity to each other, we discuss Australia and New Zealand together, although they are not twins. Australia was settled by pioneers from Asia around 70,000 years ago, when the ocean levels were much lower than now. Even with the lower ocean levels, this settlement involved a long sea journey. When ocean levels rose after the last ice age, many of the original settlements were offshore and under water. New Zealand, in contrast, was settled by seafaring people only 1500 years ago. Europeans first arrived in this area in the sixteenth century, but actual settlement by Europeans did not begin until late in the eighteenth century. As in the United States, there were conflicts between the aboriginal inhabitants and the new settlers, very fierce in Australia but surprisingly mild in New Zealand. Britain proclaimed sovereignty over New Zealand in 1840 by including it in the Australian colony of New South Wales. This merger lasted only 10 months, at which time New Zealand became an independent colony. At this time a declaration of equal rights for settlers and Maoris was made; a constitution followed 12 years later. In 1850 the six Australian states gained self-government by act of Parliament, and in 1901 they united in the Commonwealth of Australia.

Comparatively little has been written about the development of mathematics in these countries, and the present account is based largely on an article (1988) by Garry J. Tee, a professor of computer science at the University of Auckland, who has written a great deal on the history of mathematics in general. Tee says that the indigenous peoples of this area had a well-developed system of numeration and makes the point that "the common assertions to the effect that 'Aborigines have only one, two, many' derive mostly from reports by nineteenth century Christian missionaries, who commonly understood less mathematics than did the people on whom they were reporting." At the same time, he notes that these missionaries did teach Western-style mathematics to indigenous people.

6.1. Colonial mathematics. As in other countries, European colonists were not long in establishing universities in these new lands. Australia acquired universities at Sydney (1850), Melbourne (1853), Adelaide (1874), and Hobart (University of

¹² During the present author's years at Princeton (1963–1966) several of the graduate students in mathematics were Mexican students of Lefschetz.

Tasmania, 1890). In New Zealand universities opened at Dunedin (University of Otago, 1869), Christchurch (University of Canterbury, 1873), Auckland (1883), and Wellington (Victoria University, 1897). The New Zealand universities were from the beginning co-educational. The Australian Mathematical Society was founded in 1956 and the New Zealand Mathematical Society in 1974. Long before that, however, good mathematicians were being born and working in these two countries. The following short list is far from complete, but it does show that world-class mathematics and science have been produced in this region almost from the beginning.

The Bragg family. In 1915 two Australians, father and son, were awarded the Nobel Prize for physics. Each of them served as director of the Royal Institution in London. William Henry Bragg (1862–1942) was a professor of mathematics at the University of Adelaide from 1885 to 1908. His son William Lawrence Bragg (1890–1971) became Cavendish Professor of physics at Cambridge and director of the National Physical Laboratory.

Horatio Scott Carslaw. One of the standard texts on Fourier series, which was reprinted many times and eventually became immortalized in a Dover edition, was written by H. S. Carslaw (1870–1954), the third professor of mathematics at the University of Sydney (1903–1935). Carslaw was born in Scotland but moved to Australia in 1903 to take up the position at the University of Sydney. Besides his book on Fourier series, he also collaborated on a standard textbook on the Laplace transform and had an interest in the history of logarithms.

Thomas Gerald Room. Carslaw's successor at the University of Sydney was Thomas Gerald Room (1902–1986), a native of London who, like Carslaw, moved to Australia to take up an academic position. He is well remembered by combinatoricists for the concept of Room squares, about which he published a paper in 1955.

Ernest Rutherford. Another physicist with mathematical gifts was the New Zealander Ernest Rutherford (1871–1937), who studied at Canterbury University, worked at McGill University in Montreal (1898–1907), and eventually, working at Manchester University, performed a famous experiment that helped to determine the structure of the atom (the positively charged nucleus surrounded by electrons that is still the popular picture of atoms).

V. F. R. Jones. One of the brightest stars in the mathematical firmament at the moment is Vaughan Frederick Randal Jones (b. 1952), who graduated from the University of Auckland in 1973. From there he went to Switzerland, where he received the doctoral degree for a prize-winning dissertation. Since 1980 he has worked in the United States. He won the Fields Medal in 1990 for his groundbreaking work in knot theory. (The Jones polynomial is named after him.) This discovery came about while he was working in a seemingly unrelated area (von Neumann algebras) and had links to areas of mathematical physics (topological quantum field theories) that were studied by mathematical physicists such as the American Edward Witten (b. 1951) and the British topologist Simon Donaldson (b. 1957), both of whom also won the Fields Medal, Donaldson in 1986 and Witten alongside Jones in 1990. The Jones polynomial was described by the French journal *La recherche* in its issue of July–August 1997 as one of the 300 most important discoveries of the last three centuries.

Refugee mathematicians. Like the United States and Canada, Australia took in some prominent European mathematicians who were fleeing persecution during the Nazi era. Among them were Kurt Mahler (1903–1988), Hans Schwerdtfeger (1902–1990), George Szekeres (b. 1911), Hanna Neumann (1914–1971), and her husband Bernhard Neumann (1909–2002). In honor of the mathematical achievements of these refugees the Australian Mathematical Society sponsors a Mahler Lectureship, a George Szekeres Medal, and a B. H. Neumann Prize.

Ties provided by the British Commonwealth seem to have facilitated the careers of many of these people. Rutherford and Schwerdtfeger, for example, both worked for a time at McGill University in Montreal, besides the time they spent in New Zealand, Australia, Britain, and elsewhere.

7. The modern era

The advanced work in number theory, geometry, algebra, and calculus that began in the seventeenth century will be incorporated into the discussion of the mathematics itself beginning in Chapter 5. There are two reasons for not discussing it here. First, many of the names from this time on, such as Pascal, Descartes, Leibniz, Newton, Cauchy, Riemann, Weierstrass are probably already familiar to the reader from mathematics courses. Second, the increasing unity of the world makes it less meaningful to talk of “European mathematics” or “Chinese mathematics” or “Indian mathematics,” since in the modern era mathematicians the world over work on the same types of problems and use the same approaches to them. We shall now look at some general features of modern mathematics the world over.

Up to the nineteenth century mathematics for the most part grew as a wild plant. Although the academies of science of some of the European countries nourished mathematical talent once it was exhibited, there were no mathematical societies dedicated to producing mathematicians and promoting their work. This situation changed with the French Revolution and the founding of technical and normal schools to make education systematic. The effects of this change were momentous. The curriculum shifted its emphasis from classical learning to technology, and research and teaching became linked.

7.1. Educational institutions. At the time of the French Revolution the old universities began to be supplemented by a system of specialized institutions of higher learning. The most famous of these was the École Polytechnique, founded in 1795. A great deal of the content of modern textbooks of physics and mathematics was first worked out and set down in the lectures given at this institution. Admission to the École Polytechnique was a great honor, and only a few hundred of the brightest young scholars in France were accepted each year. This institution and several others founded during the time of the French Revolution, such as the École Normale Supérieure, produced a large number of brilliant mathematicians during the nineteenth century. Some of their research was devoted to questions of practical importance, such as cartography and canal building, but basic research into theoretical questions also flourished.

In Germany the unification of teaching and research proceeded from the other direction, as professors at reform-minded universities such as Göttingen (founded in 1737) began to undertake research along with their teaching. This model of development was present at the founding of the University of Berlin in 1809. This

educational trend was duplicated elsewhere in the world. During his Italian campaign Napoleon founded the Scuola Normale Superiore in Pisa, which reopened in 1843 after a long hiatus. In Russia a university opened along with the Petersburg Academy of Sciences in 1726, and the University of Moscow was founded a generation later (1755) with the aim of producing qualified professionals. It was not until the nineteenth century, however, that the faculty in Moscow began to engage in research. The University of Stockholm opened in 1878 with aims similar to those of the institutions just named. In Japan an office of translations was opened in the Shogunate Observatory in 1811. It was renamed the Institute for the Study of Foreign Books in 1857 and became the home of a department of Western mathematics in 1863, taking on two Dutch faculty members in 1865. By 1869 only Western mathematics was being taught, and the teaching was being done by French and British teachers.

7.2. Mathematical societies. Another aspect of the professionalization of mathematics was the founding of professional societies to supplement the activities of the mathematical sections in academies of sciences. The oldest of these is the Moscow Mathematical Society (founded in 1864). The London Mathematical Society was founded in 1866, the Japanese Mathematical Society in 1877. The American Mathematical Society (originally the New York Mathematical Society) was founded in 1888 and the Canadian Mathematical Society in 1945.

7.3. Journals. These educational institutions and professional societies also published their own research journals, such as the *Journal de l'École Polytechnique* and the *Journal de l'École Normale Supérieure*. These journals contained some of the most profound research of the nineteenth century. Other nations soon emulated the French. The German *Journal für die reine und angewandte Mathematik* was founded by August Leopold Crelle (1780–1855) in 1826. Informally, it is still called *Crelle's Journal*. The Italian *Annali di scienze matematiche e fisiche* appeared in 1850; the Moscow Mathematical Society began publishing the *Matematicheskii Sbornik* (*Mathematical Collection*) in 1866; the Swedish *Acta mathematica* was founded in 1881. By the end of the nineteenth century there were mathematical research journals in every European country, in North America, and in Japan. The first American research journal, *The American Journal of Mathematics*, was founded at Johns Hopkins University in 1881 with the British mathematician J. J. Sylvester as its principal editor, assisted by the American William Edward Story. The first issue of *The Canadian Journal of Mathematics* was dated 1949.

Questions and problems

3.1. Compare the way in which mathematicians have been supported in various societies discussed in this chapter. If you were in charge of distributing the federal budget, how high a priority would you give to various forms of pure and applied research in mathematics? What justification would you give for your decision? Would it involve a practical “payoff” in economic terms, or do you believe that the government has a responsibility to support the creation of new mathematics, without regard to its economic value?

3.2. Why is Seki Kōwa the central figure in Japanese mathematics? Are comparisons between him and his contemporary Isaac Newton justified?

3.3. What is the justification for the statement by the historian of mathematics T. Murata that Japanese mathematics was not a science but an art?

3.4. Why might Seki Kōwa and other Japanese mathematicians have wanted to keep their methods secret, and why did their students, such as Takebe Kenkō, honor this secrecy?

3.5. For what purpose was algebra developed in Japan? Was it needed for science and/or government, or was it an “impractical” liberal-arts subject?

3.6. Dante’s final stanza, quoted above, uses the problem of squaring the circle to express the sense of an intellect overwhelmed, which was inspired by his vision of heaven. What resolution does he find for the inability of his mind to grasp the vision rationally? Would such an attitude, if widely shared, affect mathematical and scientific activity in a society?

3.7. One frequently repeated story about Christopher Columbus is that he proved to a doubting public that the Earth was round. What grounds are there for believing that “the public” doubted this fact? Which people in the Middle Ages would have been likely to believe in a flat Earth? Consider also the frequently repeated story that people used to believe the stars were near the Earth. How is that story to be reconciled with Ptolemy’s assertion that it was acceptable to regard Earth as having the dimensions of a point relative to the stars?

3.8. What are the possible advantages and disadvantages of eliminating or greatly reducing the volume of journals, placing all articles on electronic files that can be downloaded from various information systems?

3.9. Mathematical research is like any other commercial commodity in the sense that people have to be paid to do it. We have mentioned the debate over taxing the entire public to support such research and asked the student to consider whether there is a national interest that justifies this taxation. A similar taxation takes place in the form of tuition payments to American universities. Some of the money is spent to provide the salaries of professors who are required to do research. Is there an educational interest in such research that justifies its increased cost to the student?

CHAPTER 4

Women Mathematicians

The subject of women mathematicians has become a major area in the history of mathematics over the past generation, naturally connected with the women's movement in general. Any history of mathematics should include a discussion of the conditions under which mathematics flourishes and the reasons why some people and cultures develop mathematics to a high degree while others do not. To give as complete a picture of the history of mathematics as possible we need to examine these conditions, and the case of women in mathematics is a very instructive example.

The author, who began studying the history of mathematics by researching the career of Sof'ya Kovalevskaya (1850–1891), has heard it objected that women mathematicians are receiving attention out of proportion to their mathematical merit while many talented male mathematicians are being neglected by historians. Such an objection is beside the point. Male mathematicians did not have to overcome the energy- and time-consuming obstacles that women faced. The justification for devoting a full chapter to women mathematicians and for making women mathematicians a separate area of study is very simple: Until recently, all women mathematicians had one thing in common, a societal expectation that they would spend most of their time ministering to the needs of their families. A direct corollary of that expectation was that a mathematical career should not be a woman's first priority and that societal institutions need not support or even recognize any striving for such a career. In fact, Barnard College once had a policy of firing women who got married on the grounds that "the College cannot afford to have women on the staff to whom the college work is secondary; the College is not willing to stamp with approval a woman to whom self-elected home duties can be secondary."¹ In other words, if a woman chooses to marry, her duties as a wife should be first priority. If they aren't, she is a bad woman and hence unfit to be on the staff; if they are, her duties at the College must be secondary, and again, she is unfit to be on the staff.

The subject "women mathematicians" could be replaced by a category having no reference to gender, as "mathematics practiced under conditions of discrimination." In that way the subject would be enlarged so as to include minorities such as Jewish mathematicians in Europe and the United States from the Middle Ages until the twentieth century and African Americans up to very recent years. To keep this chapter of manageable size, however, we confine it to women.

¹ http://cwp.library.ucla.edu/Phase2/Maltby_Margaret_Eliza@901234567.html

1. Individual achievements and obstacles to achievement

A useful periodization of the progress—and it *is* a story of progress—of women in mathematics, is as follows: (1) before 1800, a time when only the most exceptional woman in the most exceptionally fortunate circumstances could hope to achieve anything in mathematics; (2) the nineteenth century, a period when the support of society for a woman to have a career in mathematics was missing, but a very determined, financially independent woman could at least break into the world of science and mathematics; (3) the twentieth century, when the dam restraining women from mathematical achievement developed cracks and finally burst completely, leading to a flood of women that continues to swell right up to the present. We first discuss in general terms the obstacles that needed to be overcome, and then give brief biographies describing the lives and achievements of a number of prominent women mathematicians.

1.1. Obstacles to mathematical careers for women. In the United States many of the best graduate schools were all-male until the 1960s. A classmate of the author at Northwestern University, a very bright and mathematically talented young woman, mentioned in 1962 that she had written to an Ivy League school to inquire about study for the doctoral degree and had received a reply saying, “We have no place to house you.” A decade later, the women’s movement began to focus attention on the small number of women in mathematics, and the resulting investigation into causes has helped to remove some of the obstacles to women’s achievement in mathematics. Among the obstacles, the following have been identified:

Institutionalized discrimination. It required considerable time for society to realize that all-male institutions receiving government grants were discriminating against women. Indeed, the author’s classmate mentioned above, whatever she may have thought, did not complain publicly of discrimination for being rejected by an Ivy League school. Ironically, the existence of women’s colleges, which had arisen partly in response to this discrimination, was sometimes cited as proof that men’s colleges were not discriminatory. If the opportunities and facilities at the women’s colleges had been equal to those at the men’s colleges, that argument would have had merit; but they were not.

Discrimination went beyond the student body; it was, if anything, even worse among the faculty. Until the 1970s most universities and many companies had “anti-nepotism” rules that forbade the hiring of both a husband and wife. Since women mathematicians often married men who were mathematicians, marriage became a serious impediment to a career, whether or not the husband was supportive of his wife’s ambition. Karen Uhlenbeck (b. 1942) encountered this kind of discrimination and later wrote about it:

I was told that there were nepotism rules and that they could not hire me for this reason, although when I called them on this issue years later, they did not remember saying these things.

In earlier times Ivy League universities were not the only places women were not allowed to be. In the eighteenth century, they were not allowed to attend meetings of the Academy of Sciences in Paris nor (by social convention) to enter cafés. These were the two places where the best scientific minds of the time assembled

for conversation. The Marquise du Châtelet defied convention and went to cafés anyway, dressed as a man. In the nineteenth century women were not allowed into laboratories at some universities, so that Christine Ladd-Franklin (1847–1930) became a mathematics major even though she would have preferred physics. After writing a brilliant dissertation but being unable to obtain a degree, she turned to the new profession of psychology, but even there was shut out of professional life. In the twentieth century, when his colleagues were objecting to hiring Emmy Noether at Göttingen, Hilbert is reported (Dick, 1981, p. 168; Mackay, 1991, p. 117) to have ridiculed their objections, saying, “The Senate is not a locker room; why shouldn’t a woman go there?” As our narrative proceeds, the same three institutions – the University of London, Bryn Mawr College, and the University of Göttingen – will appear repeatedly, showing how few opportunities there were for women to pursue advanced studies in mathematics until quite recently.

The situation in the early twentieth century was described by the mathematician Gerhard Kowalewski (1876–1950) in his memoirs:²

At that time [1905] the first women students began to appear at the University of Bonn. They were still being met with harsh rejection on the part of distinguished professors at other universities, for example, Berlin, where Gustav Roethe, if he caught sight of women in the auditorium, simply refused to begin his lecture until they left the room.³ People were not so narrow-minded at Bonn. The women students formed a Society and arranged balls to which they invited their professors. There was a whole series of talented women mathematicians. Many of them took the state examination under my supervision: [among them was] Maria Vaerting, who later became a famous novelist and whose first novel... was based on her student days... At the same time she was working on a very difficult topic for a doctoral dissertation under my direction. In the end, however, she didn’t receive the doctorate as my student, since I was called to Prague. She then moved to Giessen, where her work was accepted by Professor Pasch. [Kowalewski, 1950, pp. 206–207]

Discouragement from family, friends, and society in general. We do not know what attitudes were faced by the very earliest women mathematicians, but from the eighteenth century on there are many documented cases of family opposition to such a career; particularly good examples are Sophie Germain and Sof’ya Kovalevskaya, both of whom had to go to extraordinary lengths to participate in the mathematical community. (Kovalevskaya was fortunate in being able eventually to win her

² Kowalewski believed himself to be distantly related to Vladimir Kovalevskii, husband of Sof’ya Kovalevskaya, but this connection has never been verified.

³ The kind of behavior exhibited by Roethe eventually disappeared, thanks in large part to the efforts of the Prussian Kultusminister Friedrich Althoff (1839–1908), who had asked Felix Klein (1854–1925) to be on the lookout for promising women students. In 1894, with Althoff’s approval, Klein took Grace Chisholm Young as his student, and the doors of Göttingen University were thereafter open to women. One of Althoff’s last acts as Kultusminister was to unify the education of boys and girls. Klein can be described as a liberal but not a radical, one who believed in equal opportunity for women and even affirmative action to recruit women; but he insisted that only women with demonstrated talent and background should be admitted to universities.

father's blessing on her career.) In addition, most women who have had both children and a career have had to invest more time in the children than men have done. This extra responsibility and a host of other societal expectations requiring time and effort on the part of women have made it more difficult for women to concentrate on their careers with the same single-mindedness that has characterized the most outstanding male mathematicians. In at least one case, that of Grace Chisholm Young (1868–1944), marriage meant a rather complete submersion of her talents for a time, with her husband (William H. Young, 1863–1942) getting all the credit for papers that were a joint effort. Such unequal partnerships, which seem terribly unfair a century later, were probably not common, but other such cases are known.⁴

Lack of role models. It cannot be a coincidence that many of the women “pioneers” in mathematics were the daughters of mathematicians or engineers. The absence of prominent women in these fields during the early days meant that many young girls thinking about their futures did not consider a career in technical areas. Most of the exceptions were in contact with mathematics and science from an early age because of the work their fathers did. The women who did choose such careers could get little advice from their male mentors as to how to deal with the special problems faced by a woman wishing a career in science. For example, Cathleen Morawetz, who was mentioned in Chapter 3, noticing how few job opportunities there were for women with doctorates in mathematics, nearly decided to choose a career in industry after getting her master's degree. It was her mentor Cecilia Krieger (1894–1974, later Cecilia Krieger Dunaij) who encouraged her to go to New York University. Such role models and encouragement were naturally present in greater degree at women's colleges.

Inappropriate teaching methods. The usefulness of women's colleges in helping women to develop their talents and ultimately overcome society's low expectations cannot be overemphasized. That girls, at least those being raised in traditional ways, needed to be taught differently from boys, is very clear from the following description of a geometry lesson given by Prince Bolkonskii to his daughter, Princess Mar'ya, in Leo Tolstoy's *War and Peace*.

Leaning on the table, the prince pushed forward a notebook full of geometrical diagrams.

“Now, young lady,” the old man began, bending over the notebook close to his daughter and putting one hand on the arm of the chair in which the princess was sitting, so that she felt herself completely surrounded by her father's pungent old-man and tobacco scent, so long familiar to her. “Now, young lady, these triangles are similar. Notice the angle abc ...”

The princess looked nervously at her father's sparkling eyes close by; blushes rose to her cheeks, and it was apparent that she didn't understand anything and was so frightened that fear was preventing her from understanding any of her father's subsequent reasoning, no matter how clear it was. Whether it was the fault

⁴ It is now well documented that Einstein's first wife made significant contributions to his 1905 paper on special relativity and deserved to be listed as a co-author. Although she never received the Nobel Prize in her own name, she did get Einstein's prize money under the terms of their divorce settlement.

of the tutor or of the pupil, the same thing happened every day: everything swam in front of the princess' eyes, she saw and heard nothing, but only sensed her father's dry, stern face next to her, was aware of his breath and his scent, and thought only of getting out of the study as soon as possible so that she could understand the problem in the spacious freedom of her own room. The old man made extraordinary efforts: noisily moving the chair he was sitting in back and forth, he struggled not to lose his temper; but nearly always lost it, shouted at her, and sometimes threw the notebook.

The princess had given an incorrect answer.

"What a stupid thing to say!" shouted the prince, shoved the notebook aside, and quickly turned away. But then he immediately got up, walked around, touched the princess' hair, and sat down again.

He came closer and continued his reasoning.

"No, no, Princess," he said, when at last the princess had taken the notebook with the assignments in it and was preparing to leave. "Mathematics is a great thing, young lady. I don't want you to be like those silly debutantes. Perseverance brings pleasure." He stroked her cheek with his hand. "The frivolity will eventually jump out of your head." [*War and Peace*, Book 1, Part 1, Chapt. 22]

The vividness of this scene shows that Tolstoy must have drawn it from real life. Even an enlightened father, such as Tolstoy's Prince Bolkonskii, who loved his daughter and wanted more for her than the frivolous life offered to most women in the Russian aristocracy, did not know how to carry out his own good intentions.

Sexual harassment. This painful topic has apparently not been much talked about in relation to mathematics specifically. Keith and Keith (2000) report that at a 1988 conference on women in mathematics and the sciences every woman present had experienced discrimination, not only "gender harassment... but more brutal sexual harassment." The harm that can be done by sexual harassment includes creating anxiety that interferes with work, discouraging women from seeking help in a professor's office, and blocking professional advancement for women who protest harassment or reject unwanted advances.

To struggle against all of these obstacles was the task of heroic individual women for many centuries, and what they achieved seems in many ways miraculous. Who would have guessed, for example, that a journal named *The Woman Inventor* was published more than a century ago?⁵ But real progress could be expected only when society as a whole undertook to provide support. To overcome these obstacles legislation was enacted at the federal level during the 1960s forbidding discrimination on the basis of gender. To overcome the more entrenched and subtle problems of societal discouragement and lack of role models a variety of measures have been introduced, including special workshops and institutes devoted to introducing women to mathematical research and the founding of the Association for Women in Mathematics in 1971. All major universities and corporations

⁵ It was published by Charlotte Smith (1840-1917) and managed only two issues, in April and June of 1891 (Stanley, 1992).

now have procedures for preventing and prosecuting sexual harassment. Although it cannot be said that all of these obstacles have been overcome, it is certainly the case that more and more women are choosing careers in mathematics. In many universities the number of undergraduate women majoring in mathematics is now larger than the number of men, and the number of women graduate students is approaching equality with the number of men. Equality of numbers, however, is not necessarily the goal. It may be that, given equal opportunity, more women than men would choose to be mathematicians; or the number of women freely choosing such a career might be less. What is (in the author's view) the ultimate goal—that each person should be aware of the opportunities for any career and accorded equal opportunity to pursue the career of her or his choice—has not quite been achieved, but it is fair to say that a woman can now pursue a career in mathematics and science with the same expectation of success, depending on her talent, as in any other major.

2. Ancient women mathematicians

Very few women mathematicians are known by name from early times. However, Closs (1992, p. 12) mentions a Maya ceramic with a picture of a female scribe/mathematician. From ancient Greece and the Hellenistic culture, at least two women are mentioned by name. Diogenes Laertius, in his work *Lives of Eminent Philosophers*, devotes a full chapter to the life of Pythagoras, and gives the names of his wife, daughter, and son. Since it is known that the Pythagoreans admitted women to their councils, it seems that Pythagoras' wife and daughter engaged in mathematical research at the highest levels of their day. However, nothing at all is known about any works they may have produced. All that we know about them is contained in the following paragraph from Diogenes Laertius:

Pythagoras had a wife named Theano. She was the daughter of Brontinus of Croton, although some say that she was Brontinus' wife and Pythagoras' pupil. He also had a daughter named Damo, as Lysis mentions in a letter to Hipparchus. In this letter he speaks of Pythagoras as follows: "And many say that you [Hipparchus] give public lectures on philosophy, as Pythagoras once did. He entrusted his *Commentaries* to Damo, his daughter, and told her not divulge them to anyone not of their household. And she refused to part with them, even though she could have sold them for a considerable amount of money; for, despite being a woman,⁶ she considered poverty and obedience to her father's instructions to be worth more than gold." He also had a son named Telauges, who succeeded him as head of the school, and who, according to some authors, was the teacher of Empedocles. Hippobotus, for one, reports that Empedocles described him as "Telauges, the noble youth, whom in due time, Theano bore to the sage Pythagoras." But no books by Telauges survive, although there are still some that are attributed to his mother Theano.

⁶ It is hardly worth pointing out the slur on women's character implicit in this phrase.

Hypatia. There are two primary sources for information about the life of Hypatia. One is a passage in a seven-book history of the Christian Church written by Socrates Scholasticus, who was a contemporary of Hypatia but lived in Constantinople; the other is an article in the *Suda*, an encyclopedia compiled at the end of the tenth century, some five centuries after Hypatia.⁷ In addition, several letters of Synesius, bishop of Ptolemais (in what is now Libya), who was a disciple of Hypatia, were written to her or mention her, always in terms of high respect. In one letter he requests her, being in the "big city," to procure him a scientific instrument (hygrometer) not available in the less urban area where he lived. In another he asks her judgment on whether to publish two books that he had written, saying

If you decree that I ought to publish my book, I will dedicate it to orators and philosophers together. The first it will please, and to the other it will be useful, provided of course that it is not rejected by you, who are really able to pass judgment. If it does not seem to you worthy of Greek ears, if, like Aristotle, you prize truth more than friendship, a close and profound darkness will overshadow it, and mankind will never hear it mentioned. [Fitzgerald, 1926]

The account of Hypatia's life written by Socrates Scholasticus occupies Chapter 15 of Book 7 of his *Ecclesiastical History*. Socrates Scholasticus describes Hypatia as the pre-eminent philosopher of Alexandria in her own time and a pillar of Alexandrian society, who entertained the elite of the city in her home. Among that elite was the Roman procurator Orestes. There was considerable strife at the time among Christians, Jews, and pagans in Alexandria; Cyril, the bishop of Alexandria, was apparently in conflict with Orestes. According to Socrates, a rumor was spread that Hypatia prevented Orestes from being reconciled with Cyril. This rumor caused some of the more volatile members of the Christian community to seize Hypatia and murder her in March of 415.

The *Suda* devotes a long article to Hypatia, repeating in essence what was related by Socrates Scholasticus. It says, however, that Hypatia was the wife of the philosopher Isodoros, which is definitely not the case, since Isodoros lived at a later time. The *Suda* assigns the blame for her death to Cyril himself.

Yet another eight centuries passed, and Edward Gibbon came to write the story in his *Decline and Fall of the Roman Empire* (Chapter XLVII). In Gibbon's version Cyril's responsibility for the death of Hypatia is reported as fact, and the murder itself is described with certain gory details for which there is no factual basis. (The version given by Socrates Scholasticus is revolting enough and did not need the additional horror invented by Gibbon.)

A fictionalized version of Hypatia's life can be found in a nineteenth-century novel by Charles Kingsley, bearing the title *Hypatia, or New Foes with an Old Face*. What facts are known were organized into an article by Michael Deakin (1994) and a study of her life by Maria Dzielska (1995).

3. Modern European women

Women first began to break into the intellectual world of modern Europe in the eighteenth century, mingling with the educated society of their communities, but not allowed to attend the meetings of scientific societies. The eighteenth century

⁷ This work bears the traditional name *Suidas*, erroneously thought to be the name of the person who compiled it.

produced three notable women mathematicians, whose biographies exhibit some noticeable similarities and some equally noticeable differences.

3.1. Continental mathematicians. The first two of three prominent eighteenth-century women mathematicians were the Marquise du Châtelet and Maria Gaetana Agnesi. Both were given strong classical educations at the insistence of their fathers, both took a strong interest in science, and both wrote expository works that incorporated their own original ideas. Apart from those similarities, however, there are a great many differences between the two women, beyond the obvious fact that the Marquise du Châtelet was French and Maria Gaetana Agnesi was Italian.

The Marquise du Châtelet. The Marquise du Châtelet was born Gabrielle-Émilie Tonnelier de Breteuil, the daughter of a court official of the “Sun King” Louis XIV, in 1706. She was presented at court at age 16, married to a nobleman at 19, and had a number of lovers throughout her life. She bore several children and died in 1749, apparently of complications from the birth of a child; she was 42 years old at the time.

In a preface to her translation and reworking of an English book entitled *The Fable of the Bees*, she wrote eloquently about the situation of women in general, and the difficulties she herself faced, saying

I am convinced that many women are either unaware of their talents by reason of the fault in their education or that they bury them on account of prejudice for want of intellectual courage. My own experience confirms this. [Ehrman, 1986, p. 61]

As a teenager Gabrielle-Émilie received encouragement to study mathematics from a family friend, M. de Mézières, but would have had contact with science in any case, just from being in a home where intellectual questions were taken seriously. Her scientific interests were in the area known as natural philosophy, which was the physics and chemistry of the time, but contained strong admixtures of philosophical doctrines that have since been purged. In 1740 she published *Institutions de physique*, in which she attempted a synthesis of the ideas of Newton, Descartes, and Leibniz. Five years later, she began the work for which she is best remembered, a French translation, with commentary, of Newton’s *Philosophiæ naturalis principia mathematica*. This work was published in 1756, seven years after her death.

Maria Gaetana Agnesi. In contrast to the Marquise du Châtelet, Maria Gaetana Agnesi much preferred a simple, spartan life, even though her father was the heir to a fortune made in the silk trade. Born in 1718 in Bologna, which at the time was located in the Papal States, she wanted to be a nun, and only her father’s pleading prevented her from going to a convent as a young woman. She never married and spent her time at home in activities that would be appropriate to a convent, reading religious books, praying, and studying mathematics.⁸ She was encouraged in her interest in mathematics by a monk who was also a mathematician and who frequently visited her father. In the preface to her book *Istituzioni analitiche ad uso della gioventù italiana*, she expressed her gratitude for this support, saying

⁸ If those three activities seem incongruous, one should keep in mind that a considerable portion of the women mathematicians in the United States during the 1930s were nuns.

that, despite her strong interest in mathematics, she would have gotten lost without his instruction. Using the famous mathematician Jacopo Riccati (1676–1754), for whom the Riccati equation is named, as an editor, she worked methodically on this textbook for many years. Riccati even gave her some of his own results on integration. The work was published in two volumes in 1748 and 1749 and immediately recognized as a masterpiece of organization and exposition, earning praise from the Paris Academy of Sciences. The Pope at the time, Benedict XIV, had an interest in mathematics, and he appointed her to a position as reader at the University of Bologna. Soon afterward, the Academy of Bologna offered her the chair of mathematics at the university, and the Pope confirmed this offer.

However, she does not seem to have accepted the offer. Her name remained on the rolls at the university, but she devoted herself to her charitable work, with ever more zeal after her father died in 1752. She gave away her fortune to the poor and died in poverty in 1799.

As is often the case with people who are kept away from full participation in scientific circles, the originality of Maria Agnesi's work is in the organization of the material. The small part of it that has immortalized her name is a curve that she called *la versiera*, meaning the *twisted curve*. It was translated into English by her contemporary John Colson, but the translation was not published until 1801. Colson apparently confused *la versiera* with *l'avversiera*, which means *wife of the devil*. Accordingly he gave this curve the name *witch of Agnesi*, a name that has unfortunately stuck to it and is both sad and ironic, considering the exemplary character of its author.⁹

Sophie Germain. Even though she was born much later than Maria Gaetana Agnesi and the Marquise du Châtelet, the third prominent woman mathematician of the eighteenth and early nineteenth centuries, Marie-Sophie Germain, was more isolated from the intellectual world than her two predecessors. She was born in Paris during the reign of Louis XVI, on April 1, 1776. Like Maria Gaetana Agnesi, her family had grown wealthy in the silk trade, and the family home was a center of intellectual activity. She, however, was strongly discouraged from scientific studies by her family and had to stay up late and study the works of Newton and Euler (1707–1783), teaching herself Latin in order to do so. Her persistence finally won acceptance, and she was allowed to remain unmarried and devoted to her studies. Even so, those studies were not easy to conduct. Even after the French Revolution, she was not allowed to attend school. She did venture to send some of her work to Joseph-Louis Lagrange (1736–1813) under the pseudonym “M. LeBlanc,” work he found sufficiently impressive to seek her out. He was her only mentor, but the relationship between them was not nearly so close as that between Sof'ya Kovalevskaya and her adviser Weierstrass 80 years later. She conducted a famous correspondence with Adrien-Marie Legendre (1752–1833) on problems of number theory, some of which he included in the second edition of his treatise on the subject. Later she corresponded with Carl Friedrich Wilhelm Gauss (1777–1855), again disguised as “M. LeBlanc.” Although they shared a love for number theory, the two never met face to face. Sophie Germain proved a special case of Fermat's last theorem, which asserts that there are no nonzero integer solutions of $a^n + b^n = c^n$ when $n > 2$. Her special case assumes that the prime number n does not divide a , b , or c and

⁹ Despite the widely recognized name *witch of Agnesi*, Agnesi was not the first person to study this curve.

is less than 100.¹⁰ Gauss also praised her work very highly. He did not learn her identity until 1806, when French troops occupied his homeland of Braunschweig. Remembering the death of Archimedes, Sophie Germain wrote to some friends, asking them to take care that Gauss came to no harm. Gauss' opinion of her, expressed in a letter to her the following year, is often quoted:

But how to describe to you my admiration and astonishment at seeing my esteemed correspondent Monsieur LeBlanc metamorphose himself into this illustrious personage who gives such a brilliant example of what I would find it difficult to believe. The enchanting charms of this sublime science reveal themselves only to those who have the courage to go deeply into it. But when a woman, who because of her sex and our prejudices encounters infinitely more obstacles than a man in familiarizing herself with complicated problems, succeeds nevertheless in surmounting these obstacles and penetrating the most obscure parts of them, without doubt she must have the noblest courage, quite extraordinary talents and superior genius.

Remembering that Sophie Germain was completely self-taught in mathematics and had little time to learn physics, which was increasingly developing its own considerable body of theory, we can only marvel that she had the courage to enter a prize competition in 1811 for the best paper on the vibration of an elastic plate. She had to start from zero in this enterprise, and Lagrange had warned that the necessary mathematics simply did not yet exist. According to Dahan-Dalmédico (1987, p. 351), she had learned mechanics from Lagrange's treatise and from some papers of Euler "painfully translated" from Latin. It is not surprising that her paper contained errors and that she did not win the prize. Actually, no one did; she was the only one who ventured to enter. Even so, her paper contained valuable insights, in the form of modeling assumptions that allowed the aged Lagrange to derive the correct differential equations for the displacement of the middle plane of the plate. She then set to work on these equations and in 1816 was awarded a prize for her work. This work became fundamental in the development of the theory of elasticity during the nineteenth century.

Perhaps because of the inevitable deficiencies resulting from her inadequate education, but more likely because she was a woman, Sophie Germain never received the respect she obviously deserved from the French Academy of her time. Prominent academicians seem to have given her papers the minimum possible attention. In their defense, it should be said that they were a galaxy of brilliant stars—Cauchy, Poisson, Fourier, and others—and it is unfortunate that their occasional neglect of geniuses such as Sophie Germain and Niels Henrik Abel stand out so prominently.

Like the Marquise du Châtelet, Sophie Germain had a strong interest in philosophy and published her own philosophical works. She continued to work in mathematics right up to the end of her life, writing papers on number theory and

¹⁰ The divisibility hypothesis makes for a nice theorem, since it is obviously impossible to satisfy if $n = 1$ or $n = 2$. It seems to explain why those cases are exceptions. However, we now know that it is not a necessary hypothesis.

the curvature of surfaces (another interest of Gauss, but also connected with elasticity through a principle Sophie Germain had derived from Euler's work) from the time she was stricken with breast cancer in 1829 until her death in 1831.

3.2. Nineteenth-century British women. In Britain, as on the Continent, the admission of women to universities began at a very slow pace in the late nineteenth century. Before that time women had to have some means of support for private study or otherwise blaze their own trails through the wilderness. As on the Continent, the earliest women were not specialists in mathematics but had general philosophical interests.

Mary Somerville. The work of Mary Somerville, coming about 75 years later than that of the Marquise du Châtelet, bears many resemblances to the latter, being largely expository and philosophical in nature. Mary Somerville was born in Jedburgh, Scotland, on December 26, 1780, to the family of William George Fairfax, a naval officer. Like Sophie Germain, she received no encouragement toward a scientific education. Indeed, although her mother taught her to read, she had to learn to write all by herself. Education was reserved for her brothers, although she did spend one year, which she hated, in a boarding school for girls. Like Sophie Germain, she decided to educate herself. With the encouragement of an uncle, she began learning Latin, so that alongside the education given to most girls in her day—piano, painting, needlework—she was undertaking technical subjects. A chance remark of her painting tutor, overheard by Mary, to the effect that Euclid was both the secret of perspective in art and the foundation of many other sciences, led her to study geometry with her younger brother's tutor. A brief first marriage, to a naval officer who had no appreciation of her ability or her desire to learn, led to the birth of two sons. When her husband died after only three years, she returned to Scotland with her sons, where she found a circle of sympathetic friends, including the geometer John Playfair (1748–1819), editor of a famous edition of Euclid and the man who formulated the now-common version of Euclid's fifth postulate, which is known as *Playfair's Axiom*. For one solution to a mathematical problem set in a popular journal, she received a silver medal in 1811 (which, it will be remembered, was the year in which Sophie Germain unsuccessfully sought a prize from the Paris Academy for work on a much more substantial problem).

The following year she married William Somerville, an inspector of hospitals. William proved to be much more supportive than her first husband, and together they studied geology. When he was appointed as inspector to the Army Medical Board in 1816 and elected to the Royal Society, they moved to London, where they made the acquaintance of the leading scientists of the day. In an 1826 treatise on electromagnetism by Harvard professor John Farrar (1779–1853), used widely throughout American universities in the 1830s, Mary Somerville is mentioned as having performed a vanguard experiment in electromagnetic theory. In Italy it had been discovered that when a beam of violet light was used to stroke a metal needle repeatedly in the same direction for a long time, the needle became magnetized. At the time physicists speculated that this effect might be due to the particular properties of sunlight in Italy. By verifying that the same effect could be obtained in Edinburgh, Mary Somerville showed that the explanation had to be in the physics of violet light itself. Her paper on this subject was reported in a paper bearing the title "The magnetic properties of the violet rays of the solar spectrum," published

in the *Proceedings of the Royal Society* in 1826 and, as mentioned, quoted by John Farrar and thereby made famous throughout the United States.

In an interesting reciprocity with the French translation of Newton's *Principia* made by the Marquise du Châtelet, Mary Somerville made a translation of Laplace's *Mécanique céleste* (as, it will be recalled, Nathaniel Bowditch had also done). Like the Marquise and Bowditch, Mary Somerville went far beyond merely translating, supplementing Laplace's laconic style with extensive commentaries. This work was published in 1831 and was a great success. Her book *The Connection of the Physical Sciences* (1834) went through many editions, and its speculation on the existence of an eighth planet, eventually to be known as Neptune, beyond Uranus (which had been discovered in 1781), inspired one of the co-discoverers of that planet. According to Baker (1948), her next book, *Physical Geography* (1848), was less successful from her own point of view, although not from the point of view of the experts. She was disappointed that it went through only six editions, and blamed its lack of commercial success on the appearance of cheap imitations that were "just keeping within the letter of the law [on plagiarism]." She began it in 1839 but had to delay because of her husband's illness and the need to revise her earlier book. Then, just when the manuscript was ready to go to press, another book on geography, entitled *Cosmos* and written by the great German scholar Alexander von Humboldt (1769–1859) appeared, apparently discouraging her so greatly that she considered burning her manuscript and had to be persuaded by friends to allow it to be published. It finally appeared in 1848. The reviews from those capable of reading it were glowing. Humboldt himself wrote to her, "I do not know of any book on geography in any language that can be compared with yours. You have not missed any fact or any of the grand sights of nature," and he signed himself "the author of the imprudent *Cosmos*." The subject of geography was not yet established in the curriculum in Britain, and her two-volume work did much to gain it a secure place.

As a result of these and other works, Mary Somerville was elected to a number of professional societies, including the Société de Physique et d'Histoire Naturelle in Geneva (1834), the Royal Irish Academy (1834), the Royal Astronomical Society (1835), the American Geographical and Statistical Society (1857), and the Italian Geographical Society (1870). She also won a number of academic honors. Recognizing the need for women to be liberated from their traditional confinement to the home, Mary Somerville was the first to sign the petition to Parliament organized by the philosopher John Stuart Mill (1806–1873), asking that the right to vote be extended to women. (Together with his wife, Mill had written a book entitled *The Enfranchisement of Women* and had also published Mary Wollstonecraft's *The Subjugation of Women*.) As in the United States, this right was finally granted just after the end of World War I. In 1862 she petitioned the University of London on behalf of women seeking degrees. (Note that she was 82 years old at the time!) Although this petition was rejected at that time, the University was awarding degrees to women only a few years later. Her long life finally came to an end near the end of her ninety-second year, on November 29, 1872, in Naples, Italy, where much of her geographical research had been done.

Florence Nightingale. Occasionally, new mathematics is created when people who are not professional mathematicians exercise their mathematical imaginations to

solve urgent practical problems. Most often today such work comes from physicists, who state mathematical conjectures based on their physical intuition; the conjectures are then either proved or modified by other mathematicians or mathematical physicists. The most useful mathematics from a social point of view is the mathematics used every day to settle important questions. Today, that generally means statistics. We are used to seeing histograms, line graphs, and pie charts in our newspapers, and most professional journals of even moderate technical pretensions will have articles referring to standard deviations, chi-square tests, p -numbers, and related concepts. The graphical representations of data that we are used to seeing in our newspapers owe something to the imagination of this remarkable woman.

She was born in Italy on May 12, 1820, the second daughter of a wealthy couple who were taking an extended trip. She was about one year younger than Victoria, heir to the British throne. As happened with many women of achievement, Florence's father took an interest in the education of his daughters and both encouraged and tutored them. Her decision to enter the health professions, taken in 1837, the year that Queen Victoria came to the throne, was made, she later said, as the result of a direct (though nonspecific) call from God. By the late 1840s she had persuaded her family to allow her to travel on the Continent and study the operation of hospitals. She had less technical training and inclination than did Maria Gaetana Agnesi, but she was able to integrate her technical competence with the charitable and public health activity that was her primary occupation.

The central episode in the life of Florence Nightingale was the Crimean War of 1854-1855, in which Britain and France compelled Russia to remove its fleet and fortifications from the Black Sea. Deaths from battle in this war, which was essentially a siege of the fortress of Sevastopol, were fewer than deaths from disease. Florence Nightingale was appointed to lead a party of 38 nurses to the front to treat wounded soldiers. Seeing the conditions that existed there, she was inspired to write, in collaboration with William Farr (1807-1883), a series of papers on public health, complete with statistics on the numbers and cause of deaths, which were presented in the form of a polar diagram, an early version of what we now recognize as a pie chart (Plate 5). In 1860, for this and other such innovations in data handling, she became the first woman elected a fellow of the Statistical Society. Because of her dedication to caring for the sick, comparisons with Maria Gaetana Agnesi naturally come to mind. One important difference between the two women appears to be Florence Nightingale's greater organizing skills and her belief in social rather than individual action. The explanation probably lies in the fact that the two women were born a century apart and that Florence Nightingale lived in a society where people felt themselves to have some influence over government. Maria Gaetana Agnesi, who grew up in the artistically fruitful but politically chaotic eighteenth-century Italy, probably did not have that sense of a duty to participate in political life.

Despite being an invalid for many years before her death at age 90 in 1910, Florence Nightingale worked constantly to improve health standards. To this end she published over 200 books and pamphlets, many of which are still read and still influential today. In 1907 she became the first woman awarded the Order of Merit. A museum in London is dedicated to her life and work, and links to information about her can be found at its website:

<http://www.florence-nightingale.co.uk>

3.3. Four modern pioneers. The struggle for a woman's right to be a scientist or mathematician was very much an obstacle course, similar to running the high hurdles. The first hurdle was to get the family to support a scientific education. That hurdle alone caused many to drop out at the very beginning, leaving only a few lucky or very determined women to go on to the second hurdle, gaining access to higher education. All of the women discussed above had only private tutoring in mathematics. The second hurdle began to be surmounted in the late nineteenth century. On the continent a few women were admitted to university lectures without being matriculated, as exceptional cases. These cases established a precedent, and the exceptions eventually became regularized. In Britain the University of London began admitting women in the 1870s, and in the United States there were women's colleges for undergraduate education. The opening of Bryn Mawr College in 1885 with a program of graduate studies in mathematics was an important milestone in this progress. Once a woman had surmounted the second hurdle, the third and highest of all had to be faced: getting hired and accepted as a scientist. The four pioneers we are about to discuss had to improvise their solutions to this problem. The fundamental societal changes needed to provide women with the same assured, routine access that men enjoyed when pursuing such a career required many decades to be recognized and partially implemented.

Charlotte Angas Scott. One of the first women to benefit from the relaxation of restrictions on women's education was Charlotte Angas Scott, who was born in Lincoln, England on June 8, 1858. Like many of the earlier women, she was fortunate in that her parents encouraged her to study mathematics with a tutor. She attended Girton College, Cambridge and took the comprehensive Tripos examination at Cambridge in 1880, being ranked as the eighth Wrangler (that is, she was eighth from the top of the class of mathematics majors). However, the Tripos alone was not enough to earn her a degree at Cambridge. She was very fortunate in being able to go on to graduate work in algebraic geometry under the direction of one of the greatest nineteenth-century mathematicians, Arthur Cayley (1821–1895). She earned a first (highest-rank) degree from the University of London in 1882 and, with Cayley's recommendation, the Ph.D. in 1885. Having now surmounted the second hurdle, she faced the third and highest one: finding an academic position.

Cayley, who had spent some time a few years earlier at Johns Hopkins University in Baltimore, knew that Bryn Mawr College was opening that year. On his recommendation, Scott was hired there as a professor of mathematics. There she was able to set rigorous standards for the mathematical curriculum. When the American Mathematical Society was founded a few years later, she was a member of its first Council. Another of the nine women among the original membership of the AMS was her first Ph.D. student. Her contributions to mathematical scholarship were impressive. She published one paper giving a new proof of an important theorem of Max Noether (1844–1921) in the *Mathematische Annalen*, a very prestigious German journal, and many papers in the *American Journal of Mathematics*, which had been founded by her countryman James Joseph Sylvester (1815–1897) when he was head of mathematics at Johns Hopkins. From 1899 to 1926 she was an editor of this journal, and in 1905 she became vice-president of the American Mathematical Society.

Near the end of her career the American Mathematical Society held a conference in her honor at Bryn Mawr, and one of the speakers was the great British

philosopher-mathematician Alfred North Whitehead (1861–1947), who paid tribute to her work in promoting a community of scholars, saying, “A life’s work such as that of Professor Charlotte Angas Scott is worth more to the world than many anxious efforts of diplomatists. She is a great example of the universal brotherhood of civilisations.”

Charlotte Angas Scott retired from Bryn Mawr in 1924. The following year she returned to Cambridge, where she lived the rest of her life. She died in 1931.

Sof’ya Kovalevskaya. Most of the women discussed up to now came from a leisured class of people with independent incomes. Only such people can afford both to defy convention and to spend most of their time pursuing what interests them. However, merely having an independent income was not in itself sufficient to draw a young woman into a scientific career. In most cases, some contact with intellectual circles was present as well. Hypatia was the daughter of a distinguished scholar, and Maria Gaetana Agnesi’s father encouraged her by hiring tutors to instruct her in classical languages. In the case of Sof’ya Kovalevskaya, the urge to study mathematics and science fused with her participation in the radical political and social movements of her time, which looked to science as the engine of material progress and aimed to establish a society in accordance with the ideals of democracy and socialism.

She was born Sof’ya Vasil’evna Kryukovskaya in Moscow, where her father was an officer in the army, on January 15, 1850 (January 3 on the Julian calendar in effect in the Russia of her day). As a child she looked with admiration on her older sister Anna (1843–1887) and followed Anna’s lead into radical political and social activism. According to her Polish tutor, she showed talent for mathematics when still in her early teens. She also showed great sympathy for the cause of Polish independence during the rebellion of 1863, which was crushed by the Tsar’s troops. When she was 15, one of her neighbors, a physicist, was impressed upon discovering that she had invented the rudiments of trigonometry all by herself in order to read a book on optics; he urged her father to allow her to study more science. She was allowed to study up through the beginnings of calculus with a private tutor in Saint Petersburg, but matriculation at a Russian university did not appear to be an option. Thinking that Western Europe was more enlightened in this regard, many young Russian women used a variety of methods to travel abroad. Some were able to persuade their parents to let them go. Others had to adopt more radical means, either running away or arranging a fictitious marriage, in Sof’ya’s case to a young radical publisher named Vladimir Onufrevich Kovalevskii (1842–1883). They were married in 1868 and soon after left for Vienna and Heidelberg, where Kovalevskaya studied science and mathematics for a year without being allowed to enroll in the university, before moving on to Berlin with recommendations from her Heidelberg professors to meet the dominant influence on her professional life, Karl Weierstrass (1815–1897). At Berlin also, the university would not accept her as a regular student, but Weierstrass agreed to tutor her privately. (Comparisons with the relationship between Charlotte Angas Scott and Arthur Cayley inevitably come to mind here.)

Although the next four years were extremely stressful for a number of personal reasons, her regular meetings with Weierstrass brought her knowledge of mathematical analysis up to the level of the very best students in the world (those attending Weierstrass’ lectures). By 1874, Weierstrass thought she had done more than enough work for a degree and proposed three of her papers as dissertations.

Since Berlin would not award the degree, he wrote to the more liberal University of Göttingen and requested that the degree be granted *in absentia*. It was, and one of the three papers became a classic work in differential equations, published the following year in the most distinguished German journal, the *Journal für die reine und angewandte Mathematik*.

The next eight years may well be described as Kovalevskaya's wandering in the intellectual wilderness. She and Vladimir, who had obtained a doctorate in geology from the University of Jena, returned to Russia; but neither found an academic position commensurate with their talents. They began to invest in real estate, in the hope of gaining the independent wealth they would need to pursue their scientific interests. In 1878 Kovalevskaya gave birth to a daughter, Sof'ya Vladimirovna Kovalevskaya (1878-1951). Soon afterward, their investments failed, and they were forced to declare bankruptcy. Vladimir's life began to unravel at this point, and Kovalevskaya, knowing that she would have to depend on herself, reopened her mathematical contacts and began to attend mathematical meetings. Recognizing the gap in her résumé since her dissertation, she asked Weierstrass for a problem to work on in order to re-establish her credentials. While she was in Paris in the spring of 1883, Vladimir (back in Russia) committed suicide, leading Sof'ya to an intense depression that nearly resulted in her own death. When she recovered, she resumed work on the problem Weierstrass had given her. Meanwhile, Weierstrass and his student Gösta Mittag-Leffler (1846-1927) collaborated to find her a teaching position at the newly founded institution in Stockholm.¹¹ At first she was *Privatdozent*, meaning that she was paid a certain amount for each student she taught. After the first year, she received a regular salary. She was to spend the last eight years of her life teaching at this institution.

In the mid-1880s, Kovalevskaya made a second mathematical discovery of profound importance. Mathematical physics is made complicated by the fact that the differential equations used to describe even simple, idealized cases of physical laws are extremely difficult to solve. The obstacle consists of two parts. First, the equations must be reduced to a set of integrals to be evaluated; second, those integrals must be computed. In many important cases, such as the equations of the three-body problem, the first is impossible using only algebraic methods. When it is possible, the second is often impossible using only elementary functions. For example, the equation of pendulum motion can be reduced to an integral, but that integral involves the square root of a cubic or quartic polynomial; it is known as an *elliptic integral*. The six equations of motion for a rigid body in general cannot be reduced to integrals at all using only algebraic surfaces. In Kovalevskaya's day only two special cases were known in which such a reduction was possible, and the integrals in both cases were elliptic integrals. Only in the case of bodies satisfying the hypotheses of both of these cases simultaneously were the integrals elementary. With Weierstrass, however, Kovalevskaya had studied not merely elliptic integrals, but integrals of completely arbitrary algebraic functions. Such integrals were known as *Abelian integrals* after Niels Henrik Abel (1802-1829), the first person to make significant progress in studying them. She was not daunted by the prospect of working with such integrals, since she knew that the secret of taming them was to use the functions known as *theta functions*, which had been introduced earlier by Abel and his rival in the creation of elliptic function theory, Carl Gustav Jacobi

¹¹ It is now the University of Stockholm.



Cher Monsieur !

Je vous remercie pour Votre invitation
pour demain et je viendrais avec
plaisir.

Les équations dif. qu'il s'agit d'intégrer
sont les suivantes

$$\left\{ \begin{array}{l} A \frac{dx}{dt} = (B-C)gt + g_0 t^2 - y_0 t'' \\ B \frac{dy}{dt} = (C-A)rt + g_0 t^2 - z_0 t'' \\ C \frac{dz}{dt} = (A-B)pt + g_0 t^2 - x_0 t'' \end{array} \right. \quad \begin{array}{l} \frac{dx}{dt} = g_0 t'' - r t' \\ \frac{dy}{dt} = r t' - p t'' \\ \frac{dz}{dt} = p t' - q t'' \end{array}$$

Jusqu'à présent il n'ont été intégrés que
dans deux cas : 1) $x_0 = y_0 = z_0 = 0$ (le cas de Poisson
et de Jacobi)

$$2) A = B \quad x_0 = y_0 = 0$$

les cas de Lagrange

Moi j'ai trouvée l'intégrale aussi dans le
cas où $A = B = 2C \quad z_0 = 0$ et je puis
montrer que ces 3 cas sont les seuls où l'intégrale

First page of an undated letter from Kovalevskaya. Probably the letter was written in June 1886 and meant for Charles Hermite. The reason it was not sent is probably that she saw Hermite in person before posting it. The letter communicates her discovery of a completely integrable case of the equations of motion of a rigid body about a fixed point under the influence of gravity. Courtesy of the Institut Mittag-Leffler.

(1804–1851). All she had to do was reduce the equations of motion to integrals; evaluating them was within her power, she knew. Unfortunately, it turns out that the completely general set of such equations cannot be reduced to integrals. But Kovalevskaya found a new case, much more general than the cases already known, in which this reduction was possible. The algebraic changes of variable by which she made this reduction are quite impressive, spread over some 16 pages of one of the papers she eventually published on this subject. Still more impressive is the 80-page argument that follows to evaluate these integrals, which turn out to be hyperelliptic, involving the square root of a fifth-degree polynomial. This work so impressed the leading mathematicians of Paris that they decided the time had come to propose a contest for work in this area. When the contest was held in 1888, Kovalevskaya submitted a paper and was awarded the prize. She had finally reached the top of her profession and was rewarded with a tenured position in Stockholm. Sadly, she was not to be in that lofty position for long. In January 1891 she contracted pneumonia while returning to Stockholm from a winter vacation in Italy and died on February 10.



Bronze bust of Sof'ya Kovalevskaya, placed outside the Institut Mittag-Leffler in Djursholm, Sweden on January 15, 2000, the 150th anniversary of her birth.

Resistance from conservatives. Lest it be thought that the presence of such powerful talents as Charlotte Angas Scott and Sof'ya Kovalevskaya removed all doubt as to women's ability to create mathematics, we must point out that minds did not simply change immediately. Confronted with the evidence that good women mathematicians had already existed, the geometer Gino Loria (1862-1954) rationalized his continuing opposition to the admission of women to universities as follows, in an article in *Revue scientifique* in 1904:

As for Sophie Germain and Sonja Kowalevsky, the collaboration they obtained from first-rate mathematicians prevents us from fixing with precision their mathematical role. Nevertheless what we know allows us to put the finishing touches on a character portrait of any woman mathematician. She is always a child prodigy, who, because of her unusual aptitudes, is admired, encouraged, and strongly aided by her friends and teachers. In childhood she manages to surpass her male fellow-students; in her youth she succeeds only in equalling them; while at the end of her studies, when her comrades of the other sex are progressing vigorously and boldly, she always seeks the support of a teacher, friend, or relative; and after a few years, exhausted by efforts beyond her strength, she finally abandons a work which is bringing her no joy.

The analysis of the factual errors and statistical and logical fallacies in this farrago of nonsense is left to the reader (see Problem 4.9 below). Loria could have known better. Six years before Loria wrote these words Felix Klein was quoted by the journal *Le progrès de l'est* as saying that he found his women students to be in every respect the equals of their male colleagues.

Grace Chisholm Young. Klein began taking on women students in the 1890s. The first of these students was Grace Chisholm, who completed the doctorate under his supervision in 1895 with a dissertation on the algebraic groups of spherical trigonometry. Her life and career were documented by her daughter and written up in an article by I. Grattan-Guinness (1972), which forms the basis for the present essay.

She was born on March 15, 1868, near London, the fifth child of parents of modest but comfortable means and the third child to survive. As a child she was stricken with polio and never completely recovered the use of her right hand. Like Charlotte Angas Scott, she was tutored at home and passed the Cambridge Senior Examination in 1885. Also like Scott, she attended Girton College and met Cayley. Her impressions of him were not flattering. To her he seemed to be a lumbering intellectual dinosaur, preventing any new life from emerging to enjoy the mathematical sunshine. In a colorful phrase, she wrote, "Cayley, unconscious himself of the effect he was having on his entourage, sat, like a figure of Buddha on its pedestal, dead-weight on the mathematical school of Cambridge" (Grattan-Guinness, 1972, p. 115).

In her first year at Cambridge she might have been tutored by William Young (1863-1942), who later became her husband, except that she heard that his teaching methods were ill suited to young women. She found that Newnham College, the other women's college at Cambridge, had a much more serious professional atmosphere than Girton. She made contacts there with two other young women

who had the same tutor that she had. With the support of this tutor and her fellow women students, she began to move among the serious mathematicians at Cambridge. In particular, she made friends with a student named Isabel Maddison (1869–1950) of Newnham College, who was being tutored by William Young. It will be recalled that a decade earlier Charlotte Angas Scott had been eighth Wrangler in the Tripos. In 1890, after reading a few names of the top Wranglers, the moderator—W. W. Rouse Ball (1850–1925), the author of a best-selling popular history of mathematics—made a long pause to get the attention of the audience, then said in a loud, clear voice, “*Above the Senior Wrangler: Fawcett, Newnham.*” The young woman, Philippa Fawcett¹² of Newnham College, had scored a major triumph for women’s education, being the top mathematics student at Cambridge in her year. No better role model can be imagined for students such as Isabel Maddison and Grace Chisholm. They finished first and second respectively in the year-end examinations at Girton College the following year. That fall, due to the absence of her regular tutor, Chisholm was forced to take lessons from William Young. In 1892 she ranked between the 23rd and 24th men on the Tripos, and Isabel Maddison finished in a tie with the 27th. (The rankings went as far as 112.) As a result, each received a First in mathematics. That same year they became the first women to attempt the Final Honours examinations at Oxford, where Chisholm obtained a First and Maddison a Second. This achievement made Chisholm the first person—of either gender—to obtain a First in any subject from both Oxford and Cambridge.¹³

Unfortunately, Cambridge did not offer Grace Chisholm support for graduate study, and her application to Cornell University in the United States was rejected. As an interesting irony, then, she was forced to apply to a university with a higher standard of quality than Cornell at the time, and one that was the mathematical equal of Cambridge: the University of Göttingen. There, thanks to the liberal views of Felix Klein and Friedrich Althoff, she was accepted, along with two young American women, Mary Frances (“May”) Winston (1869–1959) and Margaret Eliza Maltby (1860–1944). In 1895, Chisholm broached the subject of a Ph. D. with Klein, who agreed to use his influence in the faculty to obtain authorization for the degree. It turned out to be necessary to go all the way to the Ministry of Culture in Berlin and obtain permission from Althoff personally. Fortunately, Althoff continued to be an enthusiastic supporter, and her final oral examination took place on April 26 of that year. She passed it and was granted the Ph.D. *magna cum laude*. She herself could hardly take in the magnitude of her achievement. More than two decades had passed since the university had awarded the Ph.D. to Sof’ya Kovalevskaya *in absentia*. Grace Chisholm had become the first woman to obtain that degree in mathematics through regular channels anywhere in Germany. She and Mary Winston were left alone together for a few minutes, which they used “to execute a war dance of triumph.” Her two companions Mary Winston and

¹² Philippa Garrett Fawcett (1868–1948) was the daughter of a professor of political economy at Cambridge. Her mother was a prominent advocate of women’s rights, and her sister was the first woman to obtain a medical degree at St. Andrew’s in Scotland. Philippa used her Cambridge education to go to the Transvaal in 1902 and help set up an educational system there. From 1905 to 1934 she was Director of Education of the London County Council.

¹³ Isabel Maddison was awarded the Bachelor of Science degree at the University of London in 1892. She received the Ph.D. at Bryn Mawr in 1896 under the supervision of Charlotte Angas Scott. She taught at Bryn Mawr until her retirement in 1926.



Portraits of Felix Klein and David Hilbert in the Mathematisches Institut and streets in Göttingen named after them.

Margaret Maltby also received the Ph. D. degree at Göttingen, Maltby (in physics) in 1895 and Winston in 1896.¹⁴

Grace Chisholm sent a copy of her dissertation to her former tutor William Young, and in the fall of 1895 they began collaboration on a book on astronomy, a project that both soon forgot in the pleasant fog of courtship. They were married in June 1896. They planned a life in which Grace would do mathematical research and William would support the family by his teaching. Grace sent off her first research paper for publication, and William, who was then 33 years old, continued tutoring. Circumstances intervened, however, to change these plans. Cambridge began to reduce the importance of coaching, and the first of their four children was born in June 1897. Because of what they regarded as the intellectual dryness of Cambridge and the need for a more substantial career for William, they moved back to Germany in the autumn of 1897. With the help of Felix Klein, William sent off his first research paper to the London Mathematical Society. It was Klein's advice a few years later that caused both Youngs to begin working in set theory. William, once started in mathematics, proved to be a prolific writer. In the words of Grattan-Guinness (1972, p. 142), he "definitely belongs to the category of creative

¹⁴ Margaret Maltby taught at Barnard College (now part of Columbia University in New York) for 31 years and was chair of physics for 20 of those years. Mary Winston had studied at Bryn Mawr with Charlotte Angas Scott. She had met Felix Klein at the World's Columbian Exposition in Chicago in 1893 and had moved to Göttingen at his invitation. After returning to the United States she taught at Kansas State Agricultural College, married Henry Newson, a professor of mathematics at the University of Kansas, bore three children, and went back to teaching after Henry's early death. From 1921 to 1942 she taught at Eureka College in Illinois.

men who published more than was good for him." Moreover, he received a great deal of collaboration from his wife that, apparently by mutual consent, was not publicly acknowledged. He himself admitted that much of his role was to lay out for Grace problems that he couldn't solve himself. To the modern eye he appears too eager to interpret this situation by saying that "we are rising *together* to new heights." As he explained in a letter to her:

The fact is that our papers ought to be published under our joint names, but if this were done neither of us get the benefit of it. No. Mine the laurels now and the knowledge. Yours the knowledge only. Everything under my name now, and later when the loaves and fishes are no more procurable in that way, everything or much under your name. [Grattan-Guinness, 1972, p. 141]

Perhaps the criticism Loria made of Sophie Germain and Sof'ya Kovalevskaya for obtaining help from first-rate mathematicians might more properly have been leveled against William Young. To the author, the rationalization in this quotation seems self-serving. Yet, the only person who could make that judgment, Grace Chisholm Young herself, never gave any hint that she felt exploited, and William was certainly a very talented mathematician in his own right, whose talent simply manifested itself very late in life.

In 1903 Cambridge University Press agreed to publish a work on set theory under both their names. That book appeared in 1906; a book on geometry appeared under both names in 1905. Grace was busy bearing children all this time (their last three children were born in 1903, 1904, and 1908) and studying medicine. She began to write mathematical papers under her own name in 1913, after William took a position in Calcutta, which of course required him to be away for long periods of time. These papers, especially her paper on the differentiability properties of completely arbitrary functions, added to her reputation and were cited in textbooks on measure theory for many decades.

Sadly, the fanaticism of World War I caused some strains between the Youngs and their old mentor Felix Klein. As a patriotic German, Klein had signed a declaration of support for the German position at the beginning of the war. Four years later, as the defeat of Germany drew near, Grace wrote to him, asking him to withdraw his signature. Of course, propaganda had been intense in all the belligerent countries during the war, and even the mildest-mannered people tended to believe what they were told and to hate the enemy. Klein replied diplomatically, saying that, "Everyone will hold to his own country in light and dark days, but we must free ourselves from passion if international cooperation such as we all desire is to assert itself again for the good of the whole" (Grattan-Guinness, 1972, p. 160). If only other scholars had been as magnanimous as Klein, German scholars might have had less justification for complaining of exclusion in the bitter postwar period. At least there was no irreparable breach between the Youngs and Klein. When Klein died in 1925, his widow thanked the Youngs for sending their sympathy, saying, "From all over the world I received such lovely letters full of affection and gratitude, so many tell me that he showed them the way on which their life was built. I had him for fifty years, this wonderful man; how privileged I am above most women..." (Grattan-Guinness, 1972, p. 171).

All four of their children eventually obtained doctoral degrees, and the pair had good grounds for being well-satisfied with their married life. When World War II

began in September 1939 they were on holiday in Switzerland, and there was real fear that Switzerland would be invaded. Grace immediately returned to England, but William stayed behind. The fall of France in 1940 enforced a long separation on them. The health of William, who was by then in his late 70s, declined rapidly, and he died in a nursing home in June 1942. Grace survived for nearly two more years, dying in England in March 1944. Grattan-Guinness (1972, p. 181) has eloquently characterized this remarkable woman:

She knew more than half a dozen languages herself, and in addition she was a good mathematician, a virtually qualified medical doctor, and in her spare time, pianist, poet, painter, author, Platonic and Elizabethan scholar—and a devoted mother to all her children. And in the blend of her rôles as scholar and as mother lay the fulfillment of her complicated personality.

Emmy Noether. Sof'ya Kovalevskaya and Grace Chisholm Young had had to improve their careers, taking advantage of the opportunities that arose from time to time. One might have thought that Amalie Emmy Noether was better situated in regard to both the number of opportunities arising and the ability to take advantage of them. After all, she came a full generation later than Kovalevskaya, the University of Göttingen had been awarding degrees to women for five years when she enrolled, and she was the eldest child of the distinguished mathematician Max Noether.¹⁵ According to Dick (1981), on whose biography of her the following account is based, she was born on March 23, 1882 in Erlangen, Germany, where her father was a professor of mathematics. She was to acquire three younger brothers in 1883, 1884, and 1889. Her childhood was quite a normal one for a girl of her day, and at the age of 18 she took the examinations for teachers of French and English, scoring very well. This achievement made her eligible to teach modern languages at women's educational institutions. However, despite the difficulties women were having at universities, as depicted by Gerhard Kowalewski, she decided to attend the University of Erlangen. There she was one of only two women in the student body of 986, and she was only an auditor, preparing simultaneously to take the graduation examinations in Nürnberg. After passing these examinations, she went to the University of Göttingen for one year, again not as a matriculated student. If it seems strange that Grace Chisholm was allowed to matriculate at Göttingen and Emmy Noether was not, the explanation seems to be precisely that Emmy Noether was a German.

In 1904 she was allowed to matriculate at Erlangen, where she wrote a dissertation under the direction of Paul Gordan (1837–1912). Gordan was a constructivist and disliked abstract proofs. According to Kowalewski (1950, p. 25) he is said to have remarked of one proof of the Hilbert basis theorem, "That is no longer mathematics; that is theology." In her dissertation Emmy Noether followed Gordan's constructivist methods; but she was later to become famous for work done from a much more abstract point of view. She received the doctorate *summa cum laude* in 1907. Thus, she surmounted the first two obstacles to a career in mathematics with only a small amount of difficulty, not much more than faced by her brother Fritz (1884–1941), who was also a mathematician. That third obstacle,

¹⁵ It will be recalled that Charlotte Angas Scott had given a new proof of a theorem by Max Noether.

however, finding work at a university, was formidable. Emmy Noether spent many years working without salary at the Mathematical Institute in Erlangen. This position enabled her to look after her father, who had been frail since he contracted polio at the age of 14. It also allowed her to continue working on mathematical ideas. For nearly two decades she corresponded with Ernst Fischer (1875–1954, Gordan's successor in Erlangen), who is best remembered for having discovered the Riesz–Fischer theorem independently of F. Riesz (1880–1956). By staying in touch with the mathematical community and giving lectures on her discoveries, she kept her name before certain influential mathematicians, namely David Hilbert (1862–1943) and Felix Klein,¹⁶ and in 1915 she was invited to work as a *Privatdozent* in Göttingen. (This was the same rank originally offered to Kovalevskaya at Stockholm in 1883.) Over the next four years Klein and Hilbert used all their influence to get her a regular appointment at Göttingen; during part of that time she lectured for Hilbert in mathematical physics. That work led her to a theorem in general relativity that was highly praised by both Hilbert and Einstein. Despite this brilliant work, however, she was not allowed to pass the *Habilitation* needed to acquire a professorship. Only after the German defeat in World War I, which was followed by the abdication of the Kaiser and a general spirit of reform in Germany, was she allowed to “habilitate.” Between Sof’ya Kovalevskaya and Emmy Noether there was a curious kind of symmetry: Kovalevskaya was probably aided in her efforts to become a student in Berlin because many of the students were away at war at the time. Noether was aided in her efforts to become a professor by an influx of returning war veterans. She began lecturing in courses offered under the name Dr. Emmy Noether (without any mention of Hilbert) in the fall of 1919. Through the efforts of Richard Courant (1888–1972) she was eventually granted a small salary for her lectures.

In the 1920s she moved into the area of abstract algebra, and it is in this area that mathematicians know her work best. Noetherian rings became a basic area of study after her work, which became part of a standard textbook by her student Bartel Leendert van der Waerden (1903–1996). He later described her influence on this work (1975, p. 32):

When I came to Göttingen in 1924, a new world opened up before me. I learned from Emmy Noether that the tools by which my questions could be handled had already been developed by Dedekind and Weber, by Hilbert, Lasker, and Macaulay, by Steinitz and by Emmy Noether herself.

Of all the women we have discussed Emmy Noether was unquestionably the most talented mathematically. Her work, both in quantity and quality, places her in the elite of twentieth-century mathematicians, and it was recognized as such during her lifetime. She became an editor of *Mathematische Annalen*, one of the two or three most prestigious journals in the world. She was invited to speak at the International Congress of Mathematicians in Bologna in 1928 and in Zürich in 1932, when she shared with Emil Artin (1898–1962) a prestigious prize for the advancement of mathematical knowledge. This recognition was clear and simple

¹⁶ Klein wrote to Hilbert, “You know that Fräulein Noether is continually advising me in my projects and that it is really through her that I have become competent in the subject.” (Dick, 1981, p. 31)



Emmy Noether



Hermann Weyl

proof of her ability, but it was still short of what she deserved. Hilbert's successor in Göttingen, Hermann Weyl (1885–1955), made this point when wrote her obituary:

When I was called permanently to Göttingen in 1930, I earnestly tried to obtain from the Ministerium a better position for her, because I was ashamed to occupy such a preferred position beside her, whom I knew to be my superior as a mathematician in many respects. I did not succeed, nor did an attempt to push through her election as a member of the Göttinger Gesellschaft der Wissenschaften. Tradition, prejudice, external considerations, weighted the balance against her scientific merits and scientific greatness, by that time denied by no one. In my Göttingen years, 1930–1933, she was without doubt the strongest center of mathematical activity there. [Dick, 1981, p. 169]

To have been recognized by one of the twentieth century's greatest mathematicians as "the strongest center of mathematical activity" at a university that was second to none in the quality of its research is high praise indeed. It is unfortunate that this recognition was beyond the capability of the Ministerium. The year 1932 was to be the summit of Noether's career. The following year, the advanced culture of Germany, which had enabled her to develop her talents to their fullest, turned its back on its own brilliant past and plunged into the nightmare of Nazism. Despite extraordinary efforts by the greatest scientists on her behalf, Noether was removed from the position that she had achieved through such a long struggle and the assistance of great mathematicians. Along with hundreds of other Jewish mathematicians, including her friends Richard Courant and Hermann Weyl (who was not Jewish, but whose wife was), she had to find a new life in a different land. She accepted a visiting professorship at Bryn Mawr, which allowed her also to lecture

at the Institute for Advanced Study in Princeton.¹⁷ Despite the gathering clouds in Germany, she returned there in 1934 to visit her brother Fritz, who was about to seek asylum in the Soviet Union. (Ironically, he was arrested in 1937, during one of the many purges conducted by Stalin, and executed as a German spy on the day the Germans occupied Smolensk in 1941.) She returned to Bryn Mawr in the spring of 1934.

Weyl, who went to Princeton in 1933, expressed his indignation at the Nazi policy of excluding “non-Aryans” from teaching. In a letter sent to Heinrich Brandt (1886–1954) in Halle he wrote:¹⁸

What impresses me most about Emmy Noether is that her research has become more and more concrete and profound. Why should this Jewess not work in the area that has led to such great achievements in the hands of the “Aryan” Dedekind? I am happy to leave it to Herrn Spengler and Bieberbach to assign mathematical modes of thought according to cultures and races. [Jentsch, 1986, p. 9]

At Bryn Mawr she was a great success and an inspiration to the women studying there. She taught several graduate and postdoctoral students who went on to successful careers, including her former assistant from Göttingen, Olga Taussky (1906–1995), who was forced to leave a tutoring position in Vienna in 1933. Her time, however, was to be very brief. She developed a tumor in 1935, but she does not seem to have been worried about its possible consequences. It was therefore a great shock to her colleagues in April 1935 when, after an operation at Bryn Mawr Hospital that seemed to offer a good prognosis, she developed complications and died within a few hours.

4. American women

In the United States higher education was open to women from the late nineteenth century on in the large, well-supported state universities. The elite eastern universities later known as the Ivy League remained mostly all-male for another century; but some of them were near women’s colleges, and some of the women from those colleges were able to take courses at places like Harvard and the University of Pennsylvania. Although mathematics in general in the United States was not yet on a par with what was being done in Europe, American women began to participate in the profession in the late nineteenth century. Our summary of this story is very incomplete, and the reader is referred to the excellent article of Green and LaDuke (1987) for complete statistics on the women mathematicians and the institutions where they studied and worked.

¹⁷ There was no chance of her lecturing at Princeton University itself, which was all-male at the time.

¹⁸ Oswald Spengler (1880–1936) was a German philosopher of history, best known for having written *Der Untergang des Abendlandes* (*The Decline of the West*). His philosophy of history, which Weyl alludes to in this quote, suited the Nazis. Although at first sympathetic to them, he was repelled by their crudity and their antisemitism. By the time Weyl wrote this letter, the Nazis had banned all mention of Spengler on German radio. Ludwig Bieberbach (1886–1982) was a mathematician of some talent who worked at Berlin during the Nazi era and edited the Party-approved journal *Deutsche Mathematik*. At the time when Weyl wrote this letter, Bieberbach was wearing a Nazi uniform to the university and enthusiastically endorsing the persecution of non-Aryans.

Christine Ladd-Franklin. The first of the two American women we shall discuss was induced by prevailing prejudice to abandon mathematics for psychology, a field in which she also encountered firm exclusion. Christine Ladd was born in New York in 1847. Her mother and aunt were advocates of women's rights. Her mother died when she was 12, and she was sent to live with her father's mother. Education for girls had come to be seen as a necessity by the American middle class, and so she was enrolled at Wesleyan Academy along with boys her age who expected to be admitted to Harvard. She herself could have no such expectations, but she did dream of attending Vassar. Her father encouraged her in her studies at Wesleyan Academy, but the grandmother she was living with was opposed to Vassar. Nevertheless, she prevailed and her mother's sister supported her financially for the first year. At Vassar she was particularly encouraged by Maria Mitchell (1818–1889, the first American woman astronomer of note). After obtaining a bachelor's degree in 1869, she spent nine years as a teacher of science and mathematics, writing articles on mathematics education that were published in England. Burnout, that familiar phenomenon among those who teach adolescents, finally set in, and she began to cast about for other careers.

Such an opportunity came along at just the right time. In 1876 Johns Hopkins University opened in Baltimore, the first American university devoted exclusively to graduate studies. Moreover, it managed to hire one of the greatest European mathematicians, James Joseph Sylvester, who, being Jewish, could not obtain a position at Cambridge or Oxford.¹⁹ By great good fortune, the name Christine Ladd was familiar to Sylvester from her articles on education. On his recommendation the university agreed to allow her to attend lectures, but only lectures by Sylvester. This restriction was lifted after the first year, and she was able to attend lectures by William Edward Story and by Charles Sanders Peirce (1839–1914, described by the British philosopher Bertrand Russell as “the greatest American thinker ever”). While working at Hopkins, she married Fabian Franklin (1853–1939), a young professor of mathematics who was born in Hungary but whose parents had moved to the United States when he was 2 years old. They were to have two children in rapid succession, one of whom died in infancy. After her marriage, she wrote her name with a hyphen as Ladd-Franklin. Under the influence of Peirce she wrote a dissertation bearing the title *The Algebra of Logic*, which was published in 1883 in the *American Journal of Mathematics*, the new journal founded by Sylvester at Story's suggestion. In fact, she published several papers in that journal, and was, by any objective standards, one of the best-qualified mathematicians in the United States. Nevertheless, Sylvester and Peirce together could not fulfill the mentoring role that Weierstrass performed for Kovalevskaya, Cayley for Charlotte Angas Scott and Klein for Grace Chisholm Young. She was unable to obtain either the Ph.D. degree or an academic position. Although she had overcome the first obstacle, getting her family's support for an education, the second and third stymied her for the rest of her life.

She had always been interested in areas of science other than mathematics, and her choice of mathematics as a major at Vassar had been partly the result of being excluded, as a woman, from the physics laboratories. In the mid-1880s she began to take an interest in psychology, especially the psychology of color perception.

¹⁹ An earlier stay at the University of Virginia in 1841, when slavery still existed, had ended in disaster for Sylvester.

She wrote a study of this subject that was published in the first volume of the *American Journal of Psychology* in 1887. Vassar awarded her an honorary doctor of laws degree that year.

The laboratories she had not been allowed to enter at Vassar were finally opened to her in Germany, where her husband took a sabbatical (in Göttingen) in 1891–1892. She took advantage of the occasion to spend some time in Berlin with the great physicist Hermann von Helmholtz (1821–1894). She presented the results of her theory and experiments at a conference in London that year.

Upon returning to the United States, she began a long quest for a degree and an academic position suited to her talents. Hopkins, where her husband continued to teach, refused her applications year after year. She continued to work independently (what else could she do?) and for many years played an active role in administering fellowships to support postdoctoral work for women. Not until 1904 was she allowed to teach a course in psychology at Hopkins. The following year her husband gave up mathematics in favor of journalism. He found a position in New York in 1910, and they moved there. Remembering the “dean’s rule” at Barnard College, no one will be surprised to learn that as a married woman, she had no hope of obtaining a position there. She was allowed, however, to lecture part-time at Columbia University during 1912–1913. In 1913 she lectured at Harvard and at Clark University, where her old professor from Johns Hopkins, William Edward Story, was chair of the Mathematics Department and the president, G. Stanley Hall (1844–1924), was a famous psychologist.²⁰ She also lectured at the University of Chicago in 1914. By this time, of course, she was no longer regarded as a mathematician; her lectures were on psychology.

Except for the position of editor for Baldwin’s *Dictionary of Philosophy and Psychology*, which she occupied from 1901 to 1905, she was excluded rather completely from participation in the professional life of a psychologist. In particular, she was not allowed to attend meetings and deliver papers. Only in 1929 was she finally able to publish a lifetime of work in psychology in her treatise *Colour and Colour Theories*. This work was published simultaneously in London and New York (which explains the British spelling of the title). In a great anticlimax in 1926, Johns Hopkins finally awarded her the Ph.D. in mathematics that she had earned 43 years earlier. One hardly knows whether the old saying “better late than never” applies in such a case. She died in 1930.

Anna Johnson Pell Wheeler. A few decades of social change can make a great deal of difference to one’s life. The social traditions and prejudice that deprived Christine Ladd-Franklin of what would have been a brilliant career were, with difficulty, overcome by one of the first American women to achieve recognition in mathematics, Anna Johnson (later Anna Johnson Pell, still later Anna Johnson Pell Wheeler). She was born in Iowa in 1883 and entered the University of South Dakota in 1899, graduating in 1903. The following year she earned a master’s degree at the University of Iowa, and then in 1905 she earned a second master’s degree at Radcliffe. She remained there another year in order to study with two of the first prominent American mathematicians, Harvard professors William Fogg Osgood

²⁰ G. Stanley Hall was the first American to obtain a doctoral degree in psychology; he had been a professor at Johns Hopkins during the early 1880s and had brought Sigmund Freud to lecture at Clark in 1910.

(1864–1943, a student of Max Noether at Erlangen) and Maxime Bôcher (1867–1918, a student of Felix Klein). Even though she was not a student at Wellesley, she was awarded a Wellesley fellowship for study abroad and went to Göttingen to attend lectures by Klein and Hilbert. Her recently widowed former professor at the University of South Dakota, Alexander Pell (1857–1921), had been corresponding with her for some time. In 1907 he came to Göttingen, and they were married. She returned with him to the United States but then went back to Göttingen to finish her doctorate. For reasons that are not clear, she did the work but did not receive the degree. Her family thought she had been pressured by her husband, who was suffering from the separation, to return to him. The only explanation she gave to the dean at Radcliffe for returning without the degree was that “in Göttingen I had some trouble with Professor Hilbert and came back to America without a degree” (Grinstein and Campbell, 1982, p. 41). She emphasized that she had written her thesis without any help from Hilbert, and so was able to submit it to the University of Chicago, where, under the supervision of another early American mathematician of distinction, Eliakim Hastings Moore (1862–1932), she was awarded the degree *magna cum laude* in 1909.

Once again, a woman had obtained a Ph.D. in mathematics at the age of only 26. Her adviser Moore sought positions for her at many universities near Chicago, where her husband was a professor at the Armour Institute of Technology. But that third-stage hurdle that has been mentioned several times before once again proved nearly insurmountable. As she wrote,

I had hoped for a position in one of the good univ. like Wisc., Ill. etc., but there is such an objection to women that they prefer a man even if he is inferior both in training and research. It seems that Professor Moore has also given up hope for he has inquired at some of the Eastern Girls' Colleges and Bryn Mawr is apparently the only one with a vacancy in Math. [Grinstein and Campbell, 1982, p. 42]

As it happened, she did not go to Bryn Mawr immediately. Her husband had a stroke that year, and she took over his teaching duties at the Armour Institute of Technology while also lecturing at the University of Chicago. She proved extremely competent at both duties. Then, from 1911 to 1918 she taught at Mount Holyoke College in Massachusetts before moving on to Bryn Mawr. When Charlotte Angas Scott retired in 1924, Anna Pell became head of the mathematics department at Bryn Mawr. Four years after the death of Alexander Pell in 1921, she married another widower, Arthur Leslie Wheeler (1871–1932, a distinguished classics scholar whose books are still being reprinted 70 years after his death). Since Wheeler had just been appointed at Princeton at the time, Anna taught only part-time at Bryn Mawr for a few years. But when he died in 1932, she went back to full-time work at Bryn Mawr and presided over the invitation to Emmy Noether that brought that distinguished mathematician there. Her own work in linear algebra allowed her to supervise the theses of many students, and was so distinguished that in 1927 she was the first woman to be invited to give a Colloquium Lecture to the American Mathematical Society.²¹ She remained at Bryn Mawr until her retirement in 1948, then continued to live in the area until her death in 1966.

²¹ The second woman was Julia Bowman Robinson – in 1980!

5. The situation today

To bring the story of the progress of women in mathematics up to the present would require writing about people whose careers are still continuing. Even when people are willing to write about themselves, reporting on what they wrote is risky; there is a danger of putting the wrong emphasis on what they have said and giving an impression that they did not intend. For that reason we shall not discuss any more biographies, but consider only how the small window of opportunity available to the pioneering women mathematicians has been enlarged to a size comparable with that available to men, and ask what more needs to be done. For examples the reader is referred to the book of Henrion (1997), which contains interviews with a number of women and is aimed at overcoming persistent stereotypes about women in the profession of mathematics.

Although a mathematical education was *formally* available to women throughout the twentieth century, social conditioning discouraged young girls from aiming at such a career. As a result, few of them ever even realized that they might have the talent to be mathematicians or scientists. Colleges of engineering and medicine were full of young men; colleges of education and nursing were full of young women. There was very little “osmosis” between these two “cells” until the women’s movement began in earnest in the 1970s. Only when significant numbers of women sought admission to scientific careers did the difficulties experienced by the few women already in those careers come to public attention. What was revealed was a wide variety of ways of discriminating—women not being admitted to some of the universities where the best work was being done, being ranked at a lower priority when applying for jobs, being asked personal questions about their families, offered lower salaries, being ignored in class, not being taken seriously in applications for graduate work, not being guided and mentored properly so as to encourage them to seek advanced degrees, being asked to do menial administrative work during probationary periods, and the like. Overcoming these problems required both antidiscrimination and affirmative-action legislation. It also involved patiently educating the public, men and women alike, in new ways of “conducting business.” University administrators had to learn how to mentor young women faculty members to channel their work into areas likely to lead to tenure. The faculty members themselves had to learn to fight against the “good citizen” impulses that got them onto too many committees, into too much curriculum work, and the sacrifice of a great deal of time trying to be the best teacher possible, all at the expense of research.²²

What now remains to be done? If we assume that the offices of affirmative action/equal opportunity at our major industries and universities are doing their job properly—and if they are not, legal redress is available for those with the courage to pursue it—the main work remaining is educational. Most of all, both boys and girls in their early teens need to be shown *how scientists actually spend their time, what their jobs consist of*. Without that kind of information, they are likely to judge a profession by the difficulty of the courses they are taking

²² This sentence was written from the point of view of what is in the best interest of a faculty member seeking tenure. The best interest of the institution and its students and the greater good of society as a whole may very well be advanced through working on committees, developing curricula, and being the best possible teacher; but unless the prevailing attitudes change at major universities, a probationary faculty member is not advised to pursue tenure through those activities.

in school. And nearly *everyone*, even a very bright student, finds mathematics difficult. Students need to be shown that a career in mathematics does not require super intelligence. What students often imagine they must do—solve some difficult, long-open problem—is definitely optional, not a necessary part of a mathematical career. This task is being addressed by mathematical organizations such as the Mathematical Association of America and the American Mathematical Society, and by various programs supported by the Department of Education and the National Science Foundation.

A secondary task is to root out the remaining stereotypes from professional mathematicians themselves. The women interviewed by Henrion for her book (1997) pointed out a number of practices within the profession that “create a ‘chilly climate’ for women both in academia in general and in mathematics in particular.” Henrion has quite astutely pointed out that the mathematical community, as a community, has its own set of expectations about how a person will work, and those expectations were set by men. How those expectations may change (or may not) as more and more women take on significant roles in the professional organizations is a development that will be interesting to observe in the future. And as the image of a set of successive obstacles that we have used above to interpret the lives of the early women mathematicians shows, the further a woman progresses, the higher the “hurdles” tend to be. Henrion (1997, p. xxxi) expressed the matter somewhat differently: “[W]omen are even further from equity the farther along in the pipeline we go.” Making professional activities gender neutral is the primary challenge for the future.

Questions and problems

4.1. In the late fourth and early fifth centuries the city of Alexandria, where Hypatia lived, was divided into Christian, Jewish, and pagan cultures. Is it merely a random event that the only woman mathematician of the time in this city with a long history of scholarship happened to come from the pagan culture?

4.2. Compare the careers of Charlotte Angas Scott and Sof'ya Kovalevskaya. In what aspects were they similar? What significant differences were there? Were these differences due to the continental circles in which Kovalevskaya moved compared to the Anglo-American milieu of Scott's career? Or were they due to individual differences between the two women?

4.3. Choose two women mathematicians, either from among those discussed in this chapter or by going to a suitable website. Read brief biographical sketches of them. Then try to match each woman with a comparable male mathematician from the same era and country. Compare their motives for studying mathematics if any motives are given, the kind of education they received, the journals where they published their work, and the kind of academic positions they occupied.

4.4. How do you account for the fact that a considerable percentage (compared to their percentage of the general population) of the women studying higher mathematics in the United States during the 1930s were Roman Catholic nuns?²³

²³ Some of these nuns produced mathematical research of high quality, for example, Sister Mary Celine Fasenmyer (1906–1996).

4.5. What were the advantages and disadvantages of marriage for a woman seeking an academic career before the twentieth century? How much of this depended on the particular choice of a husband at each stage of the career? The cases of Mary Somerville, Sof'ya Kovalevskaya, and Grace Chisholm Young will be illuminating, but it will be useful to seek more detailed sources than the narratives above.

4.6. How big a part did chance play in the careers of the early women mathematicians? (The word *chance* is used advisedly, rather than *luck*, since the opportunities that came for Sof'ya Kovalevskaya and Anna Johnson Pell Wheeler were the result of tragic misfortunes to their husbands.)

4.7. How important is (or was) encouragement from family and friends in the decision to study science? How important is it to have a mentor, an established professional in the same field, to help orient early career decisions? How important is it for a young woman to have an older woman as a role model? Try to answer these questions along a scale from "not at all important" through "somewhat important" and "very important" to "essential." Use the examples of the women whose careers are sketched above to support your rankings.

4.8. Why were most of the women who received the first doctoral degrees in mathematics at German universities foreigners? Why were there no Germans among them? In his lectures on the development of nineteenth-century mathematics (1926, Vol. 1, p. 284), Klein mentions that a 17-year-old woman named Dorothea Schlözer (apparently German, to judge by the name) had received a doctorate in economics at Göttingen a full century earlier.

4.9. How strong are the "facts" that Loria adduces in his argument against admitting women to universities? Were all the women discussed here encouraged by their families when they were young? Is it really true that it is impossible to "fix with precision" the original contributions of Sophie Germain and Sof'ya Kovalevskaya? You may wish to consult biographies of these women in which their correspondence is discussed. Would collaboration with other mathematicians make it impossible to "fix with precision" the work of any male mathematicians? Consider also the case of Charlotte Angas Scott and others. Is it true that they were exhausted after finishing their education?

Next, consider what we may call the "honor student" fallacy. Universities select the top students in high school classes for admission, so that a student who excelled the other students in high school might be able at best to equal the other students at a university. Further selections for graduate school, then for hiring at universities of various levels of prestige, then for academic honors, provide layer after layer of filtering. Except for an extremely tiny elite, those who were at the top at one stage find themselves in the middle at the next and eventually reach (what is ideally) a level commensurate with their talent. What conclusions could be justified in regard to any gender link in this universal process, based on a sample of fewer than five women? And how can Loria be sure he knows their proper level when all the women up to the time of writing were systematically locked out of the best opportunities for professional advancement? Look at the twentieth century and see what becomes of Loria's argument that women never reach the top.

Finally, examine Loria's logic in the light of the cold facts of society: A woman who wished to have a career in mathematics would naturally be well advised to find a mentor with a well-established reputation, as Charlotte Angas Scott and Sof'ya

Kovalevskaya did. A woman who did not do that would have no chance of being cited by Loria as an example, since she would never have been heard of. Is this argument not a classical example of catch-22?

4.10. Here is a policy question to consider. The primary undergraduate competition for mathematics majors is the Putnam Examination, administered the first weekend in December each year by the Mathematical Association of America. In addition to its rankings for the top teams and the top individuals, this examination also provides, for women who choose to enter, a prize for the highest-ranking woman. (The people grading the examinations do not know the identities of the entrants, and a woman can enter this competition without identifying herself to the graders.) Is this policy an important affirmative-action step to encourage talented young women in mathematical careers, or does it "send the wrong message," implying that women cannot compete with men on an equal basis in mathematics? If you consider it a good thing, how long should it be continued? Forever? If not, what criterion should be used to determine when to discontinue the separate category? Bear in mind that the number of women taking the Putnam Examination is still considerably smaller than the number of men.

4.11. Continuing the topic of the Question 4.10, what criterion should be used to determine when affirmative action policies designed to overcome the effects of past discrimination against women will have achieved their aim? For example, are these policies to be continued until 50% of all mathematics professors are women within the universities of each ranking? (The American Mathematical Society divides institutions into different rankings according to the degrees they grant; there is also a less formal but still effective ranking in terms of the prestige of institutions.) What goal is being pursued: that each man and each woman should have equal access to the profession and equal opportunity for advancement in it, or that equal numbers of men and women will choose the profession and achieve advancement? Or is the goal different from both of these? If the goal is the first of these, how will we know when it has been achieved?

The History of Mathematics: A Brief Course, Second Edition

by Roger Cooke

Copyright © 2005 John Wiley & Sons, Inc.

Part 2

Numbers

Numbers are the first association in the minds of most people when they hear the word *mathematics*. The word *arithmetic* comes into English from the Greek word *arithmós*, meaning *number*. What is nowadays called arithmetic—that is, calculation—had a different name among the Greek writers of ancient times: *logistikê*, the source of our modern word *logistics*. In the comedy *The Acharnians* by Aristophanes, the hero Dicaeopolis reflects that, arriving early for meetings of the Athenian assembly, “aporô, gráphō, paratállomai, logízomai” (“I don’t know what to do with myself; I doodle, pull my hair, and calculate”).

The different levels of sophistication in the use and study of numbers provide a convenient division into chapters for the present part of the book. We distinguish three different stages in the advancement of human knowledge about numbers: (1) the elementary stage, in which a limited set of integers and fractions is used for counting and measuring; (2) the stage of calculation, in which the common operations of addition, subtraction, multiplication, and division are introduced; and (3) the theoretical stage, in which numbers themselves become an object of interest, different kinds of numbers are distinguished, and new number systems are invented.

The first stage forms the subject matter of Chapter 5. Even when dealing with immediate problems of trade, mere counting is probably not quite sufficient numeracy for practical life: some way of comparing numbers in terms of size is needed. And when administering a more populous society, planning large public works projects, military campaigns, and the like, sophisticated ways of calculating are essential. In Chapter 6 we discuss the second stage, the methods of calculating used in different cultures.

In Chapter 7 we examine the third stage, number theory. This theory begins with the mathematicians of ancient Greece, India, and China. We look at the unique achievements of each civilization: prime, composite, triangular, square, and pentagonal numbers among the Pythagoreans, combinatorics and congruences among the Chinese and Hindus.

In Chapter 8 we discuss number systems and number theory in the modern world. Here we see how algebra led to the concept of irrational (algebraic) numbers, and the geometric representation of such numbers brought along still more (transcendental) numbers. When combined with geometry and calculus, this new algebraic view of numbers led to the theory of complex variables, which in turn made it possible to answer some very delicate questions on the relative density of prime numbers (the famous “prime number theorem”). The continuing development of these connections, as well as connections with the theory of trigonometric series, has made it possible to settle some famous conjectures: the Wiles–Taylor proof of Fermat’s last theorem and a partial proof by I. M. Vinogradov of the famous Goldbach conjecture, for example.

CHAPTER 5

Counting

Counting could conceivably occur without number words. What is required is merely a matching of the objects in two sets. The legendary American gunslinger, putting a notch in the handle of his gun for every person he has killed, is an example of such counting. A vivid example is cited by Closs (1986, p. 16) as a folk tale of the Copper Eskimo. In this story, a hunter who has killed a wolf argues with another hunter who has killed a caribou as to which animal has more hair. To decide the question, they pull out the hairs one at a time and pair them off. This mode of thought has become very familiar to mathematics students over the past century because of the rise of set theory in the undergraduate curriculum.

1. Number words

Every human language that we know about has words for numbers. In the case of languages whose long history is known—English, for example—the number words seem to be of such ancient origin that they have no obvious relation to the non-numerical words in the language. Some attempts have been made to find clues as to the origin of number words and number concepts in the grammar of various languages, but very few reliable conclusions have been reached. The guesses involved are interesting, however, and we shall look at a few of them below, taken mostly from the books of Menninger (1969) and Gow (1884).

To begin with modern English, when a person says, “I know a number of ways to prove the Pythagorean theorem,” the listener will interpret the phrase “a number of” to mean “three or more.” The word *number* here is a synonym for the fancier word *multitude* and is used to refer to any set of things having three or more members. If the speaker knew only two ways to prove the Pythagorean theorem, the word *number* would almost certainly be replaced by *couple*. This last word is one of the few collective words in English with a definite numerical meaning. It is used as a synonym for *two* when the two objects mentioned are related to each other in some way, in phrases like “I have a couple of errands to run downtown.” The connection with the number two is not quite exact here, since in very informal speech the word *couple* is often stretched to mean simply a small number. From such considerations we might form the hypothesis that one and two are instinctive concepts, and that numbers as a deliberate, conscious creation of the human mind begin with three. Support for this idea comes from the reflection that English has special words for ordinal (*second*) and partitive (*half*) concepts connected with the number two and a special word (*both*) to apply to the whole of a set of two objects, while the ordinal and partitive concepts merge for numbers three and higher (*third*, *one-third*, and so on) and the same word (*all*) is used to denote the whole of any set having more than two members. Of course, what is true of English does not always apply to other languages.

Ancient Greek and Sanskrit, as well as modern English and other European languages, share a great many root words and morphological features. In a book on Greek mathematics (1884) the British mathematician James Gow (1854-1923) of Trinity College, Cambridge speculated on the possibility of using comparative philology to discover the history of mathematical terms. He noted in particular that the words for one, two, three, and four are declinable in Greek, but not the words for five and above. That fact suggested to him that numbers above four are an artificial creation. (It also dovetails neatly with the observations of Karen Fuson, discussed in Chapter 1, on the counting abilities of children.) Gow noted that in Slavonic, which is a European language, *all* numerals are declined as feminine singular nouns (those ending in 5 or above still are, in modern Russian), but he regarded this usage as later and hence not relevant to his inquiry. He also noted that all numerals are declined in Sanskrit, but thought it an important difference that no gender could be assigned to them.

In a comprehensive study of numbers and counting (1969) the mathematician Karl Menninger (1898-1963) conjectured that the words for *one* and *two* may be connected with personal and demonstrative pronouns. In favor of Menninger's conjecture, we note that in formal writing in English, and sometimes also in formal speaking, the word *one* is used to mean an unspecified person. This English usage probably derives from similar usage of the special third-person pronoun *on* in French. Speaking of French, there is a suggestive similarity between this pronoun and the word (*un*) for one in that language. The Russian third-person pronouns (*on*, *ona*, *ono* in the three genders) are the short forms of the archaic demonstrative pronoun *onyi*, *onaya*, *onoe* (meaning *that one*), and the word for the number one is *odin*, *odna*, *odno*.

Menninger also suggested that the word *two*, or at least the word *dual*, may be related to the archaic second-person singular pronoun *thou* (*du*, still used in Menninger's native language, which is German). Menninger noted that the concept of two is closely related to the concept of "other." Consider, for example, the following sentences:

This is my favorite style of gloves. I have a second pair in my closet.

This is my favorite style of gloves. I have another pair in my closet.

Menninger suggested a connection between *three* and *through*, based on other languages, such as the Latin *tres* and *trans*. Despite these interesting connections, Menninger emphasized that the words for *cardinal* numbers have left no definite traces of their origin in the modern Indo-European languages. All the connections mentioned above could be merely coincidental. On the other hand, the *ordinal* number words *first* and *second* have a more obvious connection with non-mathematical language. The word *first* is an evolved form of *fore-est* (*foremost*), meaning the one farthest forward. The word *second* comes from the Latin word *sequor* (*I follow*).

In cultures where mathematics and counting are developed less elaborately, number words sometimes retain a direct relation with physical objects exemplifying the numbers. Nearly always, the words for numbers are also used for body parts in the corresponding number, especially fingers. In English also, we find the word *digit* used to describe both a finger or toe and the special kinds of numbers that occur in representations of the positive integers in terms of a base.

Are there languages in which body parts stand for *cardinal* numbers? Could the number two be the word for eyes, for example? Gow (1884) cited a number

of examples to show that in many languages the word for five also means *hand*, and that words for eight, nine, and ten also designate specific fingers of the hand in some languages. A survey of ways of counting around the world provides some evidence for Gow's thesis. The Bororo of Mato Grosso, for example (Closs, 1986, p. 23), use a phrase for the number five that translates literally "as many of them as my hand complete." In that same language the number seven is "my hand and another with a partner," 10 is "my fingers all together in front," and 15 is "now my foot is finished."

2. Bases for counting

Children have to be taught to count before they can talk about groups of more than four things. Beyond certain sizes, it becomes impossible for anyone to tell at a glance how many objects are present. Most people, for example, can say immediately how many letters are in a word of eight letters or fewer, but have to count for longer words. When the limit of immediate perception is reached, human ingenuity goes to work and always arrives at the idea of grouping the objects to be counted into *sets* of some definite size, then counting the number of *sets*. Thus arises the notion of a *base* for counting. It is well known and seems completely natural that in most cases this base is five or ten, the normal number of fingers on one or two hands.

2.1. Decimal systems. Decimal systems arose spontaneously in ancient Egypt, India, China, and elsewhere. The choice of 10 as a base is not itself a sign of superior wisdom. Only when combined with an efficient notation does the usefulness of a base make itself known. A place-value system greatly simplifies calculation, which is just as difficult in base 10 as in any other base when done without a place-value system.

The modern decimal system. Modern American counting—and increasingly also, British counting—has special words for thousand, million, billion,¹ then trillion (a thousand billions), quadrillion, quintillion, and so forth. Because these names change with every third decimal place, we are effectively using 1000 as a base for counting large sets. That fact shows through in the use of a comma (or period, in Europe) to separate each group of three digits from its predecessor. The largest of these units that anyone is likely to encounter in newspapers is the trillion, since it is the most convenient unit for discussing the national budget or the national debt in dollars. The Greek prefixes kilo- (thousand), mega- (million), giga- (billion), and tera- (trillion) are used to discuss the memory cells in computers, and the march of technology has made the first of these essentially negligible. The prefixes milli- (one-thousandth), micro- (one-millionth), and nano- (one-billionth) for reciprocals are used in discussing computing time. These are the units needed nowadays, and those that have names at present provide a comfortable margin around the objects to which they will be applied, so that no new units will need to be invented in the foreseeable future.

¹ A billion is a thousand millions in American and increasingly in British usage, where it originally meant a million millions.

Named powers of 10. The powers of 10 that have a name, such as hundred, thousand, and so on, vary from one society to another. Ancient Greek and modern Japanese contain special words for 10 thousand: *myrias* in Greek, *man* in Japanese. With this unit one million becomes “100 myriads” in Greek or *hyakuman* in Japanese (*hyaku* means 100). In these systems it would make more sense to insert commas every four places, rather than every three, to make reading easier. The ancient Hindus gave special names to numbers that one would think go beyond any practical use. One early poem, the *Valmiki Ramayana*, from about 500 BCE, explains the numeration system in the course of recounting the size of an army. The description uses special words for 10^7 , 10^{12} , 10^{17} , and other denominations, all the way up to 10^{55} .

2.2. Nondecimal systems. The systems still used in the United States—the last bastion of resistance to the metric system—show abundant evidence that people once counted by twos, threes, fours, sixes, and eights. In the United States, eggs and pencils, for example, are sold by the *dozen* or the *gross*. In Europe, eggs are packed in cartons of 10. Until recently, stock averages were quoted in eighths rather than tenths. Measures of length, area, and weight show other groupings. Consider the following words: *fathom* (6 feet), *foot* (12 inches), *pound* (16 ounces), *yard* (3 feet), *league* (3 miles), *furlong* (1/8 of a mile), *dram* (1/8 or 1/16 of an ounce, depending on the context), *karat* (1/24, used as a pure number to indicate the proportion of gold in an alloy),² *peck* (1/4 of a bushel), *gallon* (1/2 peck), *pint* (1/8 of a gallon), and *teaspoon* (1/3 of a tablespoon).

Even in science there remain some vestiges of nondecimal systems of measurement, inherited from the ancient Middle East. In the measurement of both angles and time, minutes and seconds represent successive divisions by 60. A day is divided into 24 hours, each of which is divided into 60 minutes, each of which is divided into 60 seconds. At that point, our division of time becomes decimal; we measure races in tenths and hundredths of a second. A similar renunciation of consistency came in the measurement of angles as soon as hand-held calculators became available. Before these calculators came into use, students (including the present author) were forced to learn how to interpolate trigonometric tables in minutes (one-sixtieth of a degree) and seconds (one-sixtieth of a minute). In physical measurements, as opposed to mathematical theory, we still divide circles into 360 equal degrees. But our hand-held calculators have banished minutes and seconds. They divide degrees decimally and of course make interpolation an obsolete skill. Since π is irrational, it seems foolish to adhere to any rational fraction of a circle as a standard unit; hand-held calculators are perfectly content to use the natural (radian) measure, and we could eliminate a useless button by abandoning the use of degrees entirely. That reform, however, is likely to require even more time than the adoption of the metric system.³

² The word is a variant of *carat*, which also means 200 milligrams when applied to the size of a diamond.

³ By abandoning another now-obsolete decimal system—the Briggsian logarithms—we could eliminate *two* buttons on the calculators. The base 10 was useful in logarithms only because it allowed the tables to omit the integer part of the logarithm. Since no one uses tables of logarithms any more, and the calculators don't care how messy a computation is, there is really no reason to do logarithms in any base except the natural one, the number e , or perhaps base 2 (in number theory). Again, don't expect this reform to be achieved in the near future.

Most peculiar of all in the English system is the common land measure, the *acre*, which is an area of 43,560 square feet.⁴ That means that a square 1-acre plot of land is $\sqrt{43560} = 66\sqrt{10} \approx 208.710$ feet on a side. The unit turns out to be convenient, in that there are exactly 640 acres in a square mile (known as a *section*), which can thus be quartered into 160-acre (quarter-section) plots, a convenient size for a farm in the American Middle West during the nineteenth century. At that time a larger unit of 36 sections (an area 6 miles by 6 miles) was called a *township*. There would thus be 144 farms in a typical township.

These examples lead to an interesting inference about the origins of practical mathematics. It seems likely that numbers were not developed as an abstract tool and then applied in particular situations. If such were the case, we would expect the same base to be used in all forms of measurement. But the distillation of a preferred base, usually 10, to be applied in all measurements, took thousands of years to arrive. Even today it is resisted fiercely in the United States, which was ironically one of the earliest countries to use a decimal system of coinage. The grouping of numbers seems to have evolved in a manner specific to each particular application, just as the English language once had specific collective nouns to refer to different groups of things: a blush of boys, a bevy of girls, a herd of cattle, a flock of sheep, a gaggle of geese, a school of fish, and others.

Bases used in other cultures. A nondecimal system reported (1937) by the American mathematicians David Eugene Smith (1860–1944) and Jekuthiel Ginsburg (1889–1957) as having been used by the Andaman of Australia illustrates how one can count up to certain limits in a purely binary system. The counting up to 10, translated into English, goes as follows: “One two, another one two, another one two, another one two, another one two. That’s all.” In saying this last phrase, the speaker would bring the two hands together. This binary counting *appears* to be very inefficient from a human point of view, but it is the system that underlies the functioning of computers, since a switch has only two positions. The binary digits or *bits*, a term that seems to be due to the American mathematician Claude Shannon (1916–2001), are generally grouped into larger sets for processing.

Although bases smaller than 10 are used for various purposes, some societies have used larger bases. Even in English, the word *score* for 20 (known to most Americans only from the first sentence of Lincoln’s Gettysburg Address) does occur. In French, counting between 60 and 100 is by 20s. Thus, 78 is *soixante dix-huit* (sixty-eighteen) and 97 is *quatre-vingt dix-sept* (four-twenty seventeen). Menninger describes a purely *vigesimal* (base 20) system used by the Ainu of Sakhalin. Underlying this system is a base 5 system and a base 10 system. Counting begins with *shi-ne* (*begin-to-be* = 1), and progresses through such numbers as *aschick-ne* (*hand* = 5), *shine-pesan* (*one away from* [10] = 9), *wan* (*both sides* = *both hands* = 10), to *hot-ne* (*whole-* [person]-*to-be* = 20). In this system 100 is *ashikne hotne* or 5 twenties; 1000, the largest number used, is *ashikne shine wan hotne* or 5 ten-twenties. There are no special words for 30, 50, 70, or 90, which are expressed in terms of the basic 20-unit. For example, 90 is *wan e ashikne hotne* (10 from 5 twenties). Counting by subtraction probably seems novel to most people, but it does occur in

⁴ The word *acre* is related to *agriculture* and comes from the Latin *ager* or Greek *agrós*, both meaning *field*.

Roman numerals ($IV = 5 - 1$), and we use subtraction to tell time in expressions such as *ten minutes to four* and *quarter to five*.⁵

3. Counting around the world








We now examine the ways used to count in a selected set of cultures in which mathematics eventually developed to the point of being written down.

3.1. Egypt. In Egypt the numbers appearing in hieroglyphics (the oldest writing) are represented as vertical strokes (|) for each individual digit, up to 9; then 10 is written as \cap , 20 as $\cap\cap$, and so on. To represent 100 the Egyptians used a symbol resembling a coil of rope. Such a system requires new symbols to be invented for higher and higher groupings, as larger and larger numbers become necessary. As the accompanying photograph shows, the Egyptians had hieroglyphic symbols for 1000 (a lotus blossom), 10,000 (a crooked thumb), 100,000 (a turbot fish), and 1,000,000 (said to be the god of the air). With this system of recording numbers, no symbol for zero was needed, nor was the order of digits of any importance, since, for example, $||| \cap \cap$ and $\cap \cap |||$ both mean 23. The disadvantage of the notation is that the symbol for each power of 10 must be written a number of times equal to the digit that we would put in its place. When hieroglyphics were invented, the Egyptians had apparently not realized that it would be useful to have names for the numbers 1 through 9, and then to name the powers of 10. Later on, in the hieratic and demotic scripts that replaced hieroglyphics, they had special symbols for 1 through 9, 10 through 90, 100 through 900, and so on, a system that was reproduced in the Greek numeration with Greek letters replacing the hieratic symbols.




























3.2. Mesopotamia. As the examples of angle and time measurement show, the successive divisions or regroupings need not have the same number of elements at every stage. The sexagesimal system appears to have been superimposed on a decimal system. In the cuneiform tablets in which these numbers are written the numbers 1 through 9 are represented by a corresponding number of wedge-shaped vertical strokes, and 10 is represented by a new symbol, a hook-shaped mark that resembles a boomerang (Fig. 1). So far we seem to have a decimal system of representation, like the Egyptian hieroglyphics. However, the next grouping is not *ten* groups of 10, but rather *six* groups of 10. Even more strikingly, the symbol for the next higher group is again a vertical stroke. Logically, this system is equivalent to a base-60 place-value system with a floating "decimal" (sexagesimal) point that the reader or writer had to keep track of mentally. Within each unit (sexagesimal rank) of this system there is a truncated decimal system that is not place-value, since the ones and tens are distinguished by different symbols rather than physical location. The number that we write as 85.25, for example, could be transcribed into this notation as 1, 25; 15, meaning $1 \cdot 60 + 25 \cdot 1 + 15 \cdot \frac{1}{60}$.

This place-value sexagesimal system goes back some 4000 years in the Middle East. However, in its original form it lacked one feature that we regard as essential today, a symbol for an empty place. The later Greek writers, such as Ptolemy in

⁵ Technology, however, is rapidly removing this last vestige of the old way of counting from everyday life. Circular clock faces have been largely replaced by linear digital displays, and ten minutes to four has become 3:50. This process began long ago when railroads first imposed standard time in place of mean solar time and brought about the first 24-hour clocks.

						
10^0	10^1	10^2	10^3	10^4	10^5	10^6

Powers of 10 from 10^0 to 10^6 in hieroglyphics.

	1	10	100
1			
2			
3			
4			
5			
6			
7			
8			
9			

Hieratic symbols, arranged as a multiplication table.

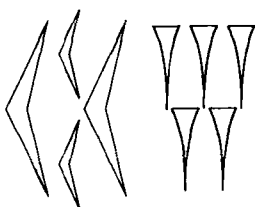


FIGURE 1. The cuneiform number 45.

the second century CE, used the sexagesimal notation with a circle to denote an empty place.

3.3. India. The modern system of numeration, in which 10 symbols are used and the value of a symbol depends only on its physical location relative to the other symbols in the representation of a number, came to the modern world from India by way of the medieval Muslim civilization. The changes that these symbols have undergone in their migration from ancient India to the modern world are shown in Fig. 2. The idea of using a symbol for an empty place was the final capstone on the creation of a system of counting and calculation that is in all essential aspects the one still in use. This step must have been taken well over 1500 years ago in India. There is some evidence, not conclusive, that symbols for an empty place were used earlier, but no such symbol occurs in the work of Aryabhata. On the other hand, such a symbol, called in Sanskrit *sunya* (empty), occurs in the work of Brahmagupta a century after Aryabhata.

3.4. China. The idea of having nine digits combined with names for the powers of 10 also occurred to the Chinese, who provided names for powers of 10 up to 100,000,000. The Chinese system of numbering is described in the *Sun Zi Suan Jing*. A certain redundancy is built into the Chinese system. To understand this redundancy, consider that in English we could write out the number 3875 in words as three thousand eight hundred seventy-five. Since Chinese uses symbols rather than letters for words, the distinction between a written number such as “seven” and the corresponding numeral 7 does not exist in Chinese. But in writing their numbers the Chinese did not use physical location as the *only* indication of the value of a digit. Rather, that value was written out in full, just as here. To convey the idea in English, we might write 3875 as 3 thousands 8 hundreds 7 tens and 5. Because of this way of writing, there is no need for a zero symbol to hold an empty place. For example, 1804 would simply be 1 thousand 8 hundreds and 4. Large numbers were handled very efficiently, with a special name for 10,000 (*wan*). Its square [*wan wan*, that is, 100,000,000 ($= 10^8$)] was called *yi*. Thereafter the Chinese had special names for each power of 10^8 . Thus, 10^{16} was *zhao*, 10^{24} was *jing*, and so on, up to 10^{80} (*zai*), which was surely large enough to meet any needs of commerce or science until the twentieth century.⁶ In that sense 10^8 amounted to a second base for arithmetic in Chinese usage.

⁶ An estimate ascribed to Sir Arthur Eddington (1882-1944) of the number of protons in the universe put the number at $136 \cdot 2^{256}$, which is approximately $1.575 \cdot 10^{79}$.

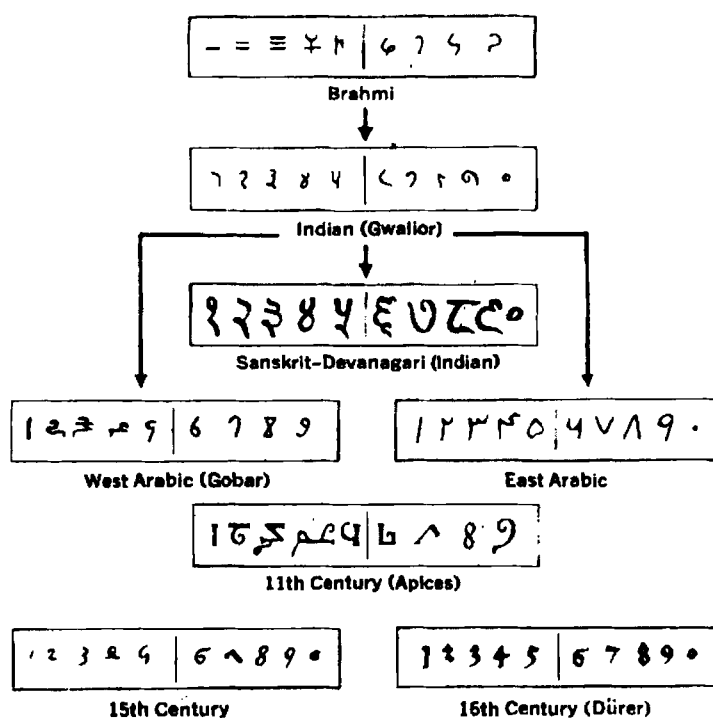


FIGURE 2. Evolution of the Hindu-Arabic numerals from India to modern Europe. ©Vandenhoeck & Ruprecht, from the book by Karl Menninger, *Zahlwort und Ziffer*, 3rd ed., Göttingen, 1979.

Some later Chinese numbering seems to reflect contact with India. Buddhism entered China during the Han dynasty (202 BCE–220 CE), and Buddhist monks had a fondness for large numbers. Unmistakable evidence of influence from India can be found in the *Suan Shu Chimeng* by Zhu Shijie, who introduced names for very large powers of 10, including the term “sand of the Ganges” for 10^{96} .

3.5. Greece and Rome. You are familiar with Roman numerals, since books still use them to number pages in the front matter, and some clock faces still show the hours in Roman numerals. Although these numerals were adequate for counting and recording, you can well imagine that they were rather inefficient for any kind of calculation. Adding, say MDCLIX to CCCIV, would take noticeably longer than adding 1659 to 304, and the idea of multiplying or dividing these two numbers seems almost too horrible to contemplate.⁷ The Greek numeral system was hardly better as far as calculation is concerned. The 24-letter Greek alphabet used today, together with 3 older letters, provided symbols for 1, ..., 9, 10, ..., 90, 100, ..., 900, essentially the system used by the Egyptians. These symbols are shown in Fig. 3. The 3 older letters were φ (digamma) for 6 (now usually written as the letter sigma in the form ς that it assumes at the end of a word), ρ (qoppa) for 90, and λ (sampi) for 900. When letters were used as numbers, they were usually given a prime, so

⁷ Nevertheless, the procedure for doing so can be learned in a fairly short time. Detlefsen and co-authors (1975) analyzed the procedure and compared it to a “paper-and-pencil abacus.”

1	2	3	4	5	6	7	8	9
α'	β'	γ'	δ'	ϵ'	ς'	ζ'	η'	θ'
10	20	30	40	50	60	70	80	90
ι'	κ'	λ'	μ'	ν'	ξ'	\omicron'	π'	ϕ'
100	200	300	400	500	600	700	800	900
ρ'	σ'	τ'	υ'	φ'	χ'	ψ'	ω'	λ'

FIGURE 3. The ancient Greek numbering system

that the number 4 would be represented by the fourth letter of the Greek alphabet (δ) and written δ' . When they reached 1000 (*khiliás*), the Greeks continued with numerical prefixes such as *tetrakishílioi* for 4000 or by prefixing a subscripted prime to indicate that the letter stood for thousands. Thus, δ' stood for 4000, and the number 5327 would be written $\epsilon'\tau'\kappa'\zeta'$. The largest independently named number in the ancient Greek language was 10,000, called *myriás*. This word is the source of the English word *myriad* and is picturesquely derived from the word for an ant (*mýrmēx*). Just how large 10,000 seemed to the ancient Greeks can be seen from the related adjective *myríos*, meaning *countless*.

The Sand-reckoner of Archimedes. The Egyptian-Greek system has the disadvantage that it requires nine new names and symbols each time a higher power of 10 is needed, and the Roman system is even worse in this regard. One would expect that mathematicians having the talent that the Greeks obviously had would realize that a better system was needed. In fact, Archimedes produced a system of numbering that was capable of expressing arbitrarily large numbers. He wrote this method down in a work called the *Psammitēs* (*Sand-reckoner*, from *psámmos*, meaning *sand*).

The problem presented as the motivation for the *Psammitēs* was a childlike question: *How many grains of sand are there?* Archimedes noted that some people thought the number was infinite, while others thought it finite but did not believe there was a number large enough to express it. That the Greeks had difficulty imagining such a number is a reflection of the system of naming numbers that they used. To put the matter succinctly, they did not yet have an awareness of the immense potential that lies in the operation of exponentiation. The solution given by Archimedes for the sand problem is one way of remedying this deficiency.

Archimedes saw that solution of the problem required a way of “naming” arbitrarily large numbers. He naturally started with the largest available unit, the myriad (10,000), and proceeded from there by multiplication and a sort of induction. He defined the first *order* of numbers to be all the numbers up to a myriad of myriads (100,000,000), which was the largest number he could make by using the available counting categories to count themselves. The second order would then consist of the numbers from that point on up to a myriad of myriads of first-order numbers, that is, all numbers up to what we would call $(10^8)^2 = 10^{16}$. The third order would then consist of all numbers beyond the second order up to a myriad of myriads of second-order numbers (10^{24}). He saw that this process could be continued up to an *order* equal to a myriad of myriads, that is, to the number $(10^8)^{10^8}$. This is a gargantuan number, a 1 followed by 800 million zeros, surely larger than any number science has ever needed or will ever need. But Archimedes realized































0 	1 	2 	3 	4 
5 	6 	7 	8 	9 
10 	11 	12 	13 	14 
15 	16 	17 	18 	19 
20 	21 	22 	23 	24 
25 	26 	27 	28 	29 

FIGURE 4. Maya numerals

the immensity of infinity. He saw that once the process just described had been completed, he could label the numbers that were named up to that point the first *period*. The mind feels carried to dizzying heights by such a process. Archimedes did not stop until he reached a number that we would represent as a 1 followed by *80 quadrillion* zeros! And of course, there is not the slightest reason to suspect that Archimedes thought the creation of integers needed to stop there. It must stop somewhere, of course.

By applying reasonable and generous estimates of the size of the universe as it was known to him, Archimedes showed that the number of grains of sand needed to fill it up could not go beyond the 1000 sixth-order units (1000 units into the seventh order, or 10^{51} in our terms). Allowing an assumption of an even larger universe, as imagined by the astronomer Aristarchus, he showed that it could not hold more than 10^{63} grains of sand.

3.6. The Maya. Although geographically far removed from Egypt, the Maya culture that existed in what is now southern Mexico and Central America from 300 BCE to 1500 CE shows some intriguing resemblances to that of ancient Egypt, especially in the building of pyramidal structures and in a hieroglyphic type of writing. On the other hand, the Maya system of counting resembles more the Mesopotamian sexagesimal system, except that it is vigesimal (base 20). As with the Mesopotamian system, only two symbols are needed to write all the numbers up to the base: a dot for ones and a horizontal line for fives. Thus the smaller base on which the vigesimal system is built is five in the case of the Maya, whereas it was 10 in the Mesopotamian system. The Maya numerals illustrate the principle that higher-level groupings need not always have the same number of members as the lower. As Fig. 4 shows, four groups of five are consolidated as a single unit of

20, and there is a cowrie-shell figure standing for zero. The second vigesimal digit in a Maya number should normally represent units of $20 \cdot 20 = 400$. However, in the *Haab* calendar, discussed below, it represents 360. The reason probably comes from the objects being counted, namely days. Even today a “business year” is 360 days (twelve 30-day months), and this cycle was also important to the Maya. Beyond this point the unit for each place value is 20 times the value of its predecessor.

4. What was counted?

People have counted an endless list of things since time immemorial. But if we were to name three items whose count was of most importance, these would be days, years, and new moons. One of the earliest uses of both arithmetic and geometry was in the construction of reliable calendars. Calendars have a practical economic value in organizing the activities of nomadic and agricultural peoples; this value is in addition to the social value associated with the scheduling of religious rites. For these reasons, calendars have been regarded as both sacred lore and applied science. At the base of any calendar must lie many years of record keeping, simply counting the days between full moons and solstices. Only after a sufficient data base has been collected can the computations needed to chart the days, weeks, months, and years be carried out. We know that such observations have been made for a long time, since the prominent lines of sight at many megalithic structures such as Stonehenge mark the summer and winter solstices. It does not require very acute observation to notice that the sun rises and sets at points farther and farther north for about 182 or 183 days, then begins to move south for the next 182 or 183 days. Once that observation was made, setting up posts to keep track of the exact location of sunrise and sunset would not have taken very long. This progression of the sun could also be correlated with the star patterns (constellations) that rise at sunset, marking the cycle we call a tropical or sidereal year. These two years actually differ by about 20 minutes, but obviously it would require a long time for that discrepancy to be noticed.

4.1. Calendars. The first broad division in calendars is between what we may call (for purposes of the present discussion only) *linear* and *cyclic* calendars. In a linear calendar the basic unit is the day, and days are simply numbered (positively or negatively) from some arbitrary day to which the number zero or 1 is assigned. Such calendars are highly artificial and used mostly for scientific purposes. For civil use, calendars attempt to repeat cycles after a month or a year or both. In traditional calendars, years were counted within the reign of a particular ruler and began with 1 as each new ruler came to power, but in the Gregorian calendar the year number does not cycle. Days and months, however, do cycle; they have their names and numbers repeated at fixed intervals. Cyclic calendars may be classified as *solar*, *lunar*, and *lunisolar*.

Ancient Egypt. The Egyptians observed the world about them with considerable accuracy, as the precise north-south orientation of some of the pyramids shows. Anyone who observes the sky for any extended period of time cannot help noticing the bright blue-green star Sirius, which is overhead at midnight during winter in the northern hemisphere. According to Montet (1974), it was recorded on the outside wall of the Temple of Ramesses III at Medinet Habu that the first day of the Egyptian year was to be the day on which Sirius and the Sun rose at the same time. To the Egyptians Sirius was the goddess Sôpdit, and they had a special

reason for noticing it. Like all stars, Sirius gains about four minutes per day on the Sun, rising a little earlier each night until finally it rises just as the Sun is setting. Then for a while it cannot be seen when rising, since the Sun is still up, but it can be seen setting, since the Sun will have gone down before it sets. It goes on setting earlier and earlier until finally it sets just after the Sun. At that point it is too close to the Sun to be seen for about two months. Then it reappears in the sky, rising just before the Sun in the early dawn. It was during these days that the Nile began its annual flood in ancient times⁸ Thus the heliacal rising of Sirius (simultaneous with the Sun) signaled the approach of the annual Nile flood. The Egyptians therefore had a very good basis for an accurate solar calendar, using the heliacal rising of Sirius as the marker of the year.

The Egyptians seem originally to have used a lunar calendar with 12 lunar cycles per year. However, such a calendar is seriously out of synchronicity with the Sun, by about 11 or 12 days per year, so that it was necessary to add an extra "intercalary" month every two or three years. All lunar calendars must do this, or else wander through the agricultural year. However, at an early date the Egyptians cut their months loose from the Moon and simply defined a month to consist of 30 days. Their calendar was thus a "civil" calendar, neither strictly lunar nor strictly solar. Each month was divided into three 10-day "weeks" and the entire system was kept from wandering from the Sun too quickly by adding five extra days at the end of the year, regarded as the birthdays of the gods Osiris, Horus, Seth, Isis, and Nephthys. This calendar is still short by about $\frac{1}{4}$ day per year, so that in 1456 years it would wander through an entire cycle of seasons. The discrepancy between the calendar and the Sun accumulated slowly enough to be adjusted for, so that no serious problems arose. In fact, this wandering has been convenient for historians, since the heliacal rising of Sirius was recorded. It was on the first day of the Egyptian year in 2773 BCE, 1317 BCE, and 139 CE. Hence a document that says the heliacal rising occurred on the sixteenth day of the fourth month of the second season of the seventh year of the reign of Senusret III makes it possible to state that Senusret III began his reign in 1878 BCE (Clayton, 1994, pp. 12–13). On the other hand, some authorities claim that the calendar was adjusted by the addition of intercalary days from time to time to keep it from wandering too far. When the Greeks came to Egypt, they used the name Sothis to refer to Sôpdit. Consequently, the period of 1456 years is known as the *Sothic cycle*.

The Julian calendar. In a solar calendar, the primary period of time being tracked is the solar year, which we now know to be 365.2422 mean solar days long. Taking 365.25 days as an approximation to this period, the Julian calendar (a solar calendar) makes an ordinary year 365 days long and apportions it out among the months, with January, March, May, July, August, October, and December each getting 31 days, while April, June, September, and November each get 30 days and February gets 28 days. For historiographical purposes this calendar has been projected back to the time before it was actually created. In that context it is called the *proleptic* Julian calendar. In a solar calendar the month is not logically necessary, and the months have only an approximate relation to the phases of the Moon.

⁸ The floods no longer occur since the Aswan Dam was built in the 1950s.

The Gregorian calendar. A modification of the Julian calendar was introduced in 1582 on the recommendation of Pope Gregory XIII. The Gregorian correction removed the extra day from years divisible by 100, but restored it for years divisible by 400. A 400-year period on the Gregorian calendar thus contains 303 years of 365 days and 97 leap years of 366 days, for a total of 146,097 days. It would be perfectly accurate if the year were of the following length:

$$365 + \frac{1}{4} - \frac{1}{100} + \frac{1}{400} = 365.2425 \text{ days.}$$

Since this figure is slightly too large, there are still too many leap years in the Gregorian calendar, and a discrepancy of one day accumulates in about 3300 years.

The Muslim calendar. The prophet Muhammed decreed that his followers should regulate their lives by a purely lunar calendar. In a lunar calendar the months are in close synchrony with the phases of the Moon, while the years need not have any close relationship to the seasons or the position of the Sun among the stars. The Muslim calendar, taking as its epoch the date of the Hijra (July 15, 622 CE), consists of 12-month years in which the odd-numbered months have 30 days and even-numbered months 29 days, except that the final month has 30 days in a leap year. Thus, the year is 354 or 355 days long, and as a result, the years wander through the tropical year.

The Hebrew calendar. More common than purely lunar calendars are lunisolar calendars, in which the months are kept in synchrony with the phases of the Moon and extra months are inserted from time to time to keep the years in synchrony with the Sun. These calendars lead to a need for *calculation* and therefore take us right up to the development of arithmetic. Several such calendars have been used since ancient times and continue to be used today in Israel, China, and elsewhere. It must have required many centuries of record keeping for the approximate equation “19 solar years = 235 lunar months” to be recognized. Since $235 = 12 \cdot 12 + 7 \cdot 13$, the addition of the extra month 7 times in 19 years will keep both years and months in balance, with an error of only about 2 hours in each 19-year cycle, or one day in 220 years.

The Julian day calendar. An example of what we have called a linear calendar is the Julian day calendar, which is to be distinguished from the Julian calendar. The Julian day calendar was invented by Joseph Justus Scaliger (1540–1609) and apparently named in honor of his father Julius Caesar Scaliger (1484–1558). It was advocated by the British astronomer John Frederick William Herschel (1792–1871). In this calendar each day is counted starting from what would be the date January 1, 4713 BCE on the Julian calendar. Thus the day on which the first draft of this paragraph was written (August 9, 2002, which is July 27, 2002 on the Julian calendar) was Julian day 2,452,496.

The Maya calendar. The most unusual calendar of all was kept by the Maya. The three Maya calendars account for a number of phenomena of astronomical and agricultural importance. As discussed above, numbers were written in a place-value system in which each unit is 20 times the next smaller unit, except that when days were being counted, the third-place unit, instead of being $20 \cdot 20 = 400$, was $20 \cdot 18 = 360$. This apparent inconsistency was probably because there are 360 days in the “regular” part of the 365-day Maya calendar known as the *Haab*, and the other 5 days were apparently regarded as unlucky (and so, best not included in the

count). This *Haab* calendar is another point of resemblance between the Maya and Egyptian civilizations.

Counting by 20 also helps to explain the rather mysterious grouping of days in a second, 260-day calendar, known as the *Tzolkin*, which is said to be from the words *tzol*, meaning *to put in order*, and *kin*, meaning *day*. One conjecture to account for this calendar is that the Maya gave a 20-day period to each of 13 gods that they worshipped. A modification of this conjecture is that the Maya formed from two different groups, one naming its days after 13 gods, the other after 20 gods, and that the *Tzolkin* made mutual comprehension easier. A second conjecture is that 260 days is approximately the growing period for maize from the time of planting to harvesting. Still a third conjecture is based on the fact that 260 days is the length of time each year in which the Sun culminates south of the zenith at a latitude of 15° N, where the southern portion of the Maya territory was located. This last explanation, however, seems inconsistent with the obvious fact that the Sun then culminates north of the zenith for the next 105 days, yet the calendar begins another 260-day cycle immediately. The two cycles coincide after 52 *Haab* years, a period called the *Calendar Round*.

An important aspect of Maya astronomy was a close observation of Venus. The Maya established that the synodic period of Venus (the time between two successive conjunctions with the Sun when Venus is moving from the evening to the morning sky) is 584 days. By coincidence, 65 synodic periods of Venus equal two Calendar Rounds (37,960 days), so that the Maya calendar bears a particularly close relation to this planet.

The third Maya calendar, the *Long Count*, resembles the Julian day calendar, in that it counts the days since its epoch, believed to be August 12, 3113 BCE on the proleptic Julian calendar. This date is not certain; it is based on the dates given in Maya inscriptions, which are presumed to be historical. Most of these dates are five-digit numbers starting with 9. Since $9.0.0.0.0 = 9 \cdot 20^2 \cdot 18 \cdot 20 = 1,296,000$ days, that is, approximately 3548 years and 4 months, and this date is associated with events believed to have occurred in 436 CE, one arrives at the stated epoch. Even though the Long Count does not explicitly mention months or years and counts only days, the Mayan place-value notation for numbers makes it possible to convert any date immediately to years, months, and days since the beginning. The digits excluding the final two represent the vigesimal (base 20) notation for a multiple of 360, while the next-to-last represents a multiple of 20, and the final digit is the number of units. Thus the Long Count date of December 31, 2000, which is 1,867,664, would be written in Maya notation with the vigesimal digits separated by commas as 12, 19, 7, 17, 4. It therefore represents 5187 *Haab* years of 360 days each ($5187 = 12 \cdot 20^2 + 19 \cdot 20 + 7$), plus 17 *Tzolkin* months of 20 days each, plus 4 days ($1,867,664 = 5187 \cdot 360 + 17 \cdot 20 + 4$). The Long Count may not be a “perpetual” day calendar. That is, it may be cyclic rather than what we have called linear. Some scholars believe that it cycles back to zero when the first of the five vigesimal digits reaches 13. Since $13.0.0.0.0 = 1,872,000$ days, or about 5125 years, the Long Count should have recycled around 1992 on the Gregorian calendar.

4.2. Weeks. The seven-day week was laid down as a basic human labor cycle in the Book of Genesis. If we look for human origins of this time period, we might associate it with the waxing and waning of the Moon, since one week is the time

required for the Moon to go halfway from full to new or vice versa. There is another, more plausible, astronomical connection, however, since there are exactly seven heavenly bodies visible to the unaided eye that move around among the fixed stars: the Sun, the Moon, and the five planets Mercury, Venus, Mars, Jupiter, and Saturn. These planets were also gods to many ancient peoples and gave their names to the weekdays in some of the Romance languages, such as French and Italian. In English the Norse gods serve the same purpose, with some identifications, such as Tiw (Tiu) with Mars, Odin (Wotan) with Mercury, Thor with Jupiter, and Frigga with Venus. Saturn was not translated; perhaps the Norse simply didn't have a god with his lugubrious reputation. This dual origin of the week, from Jewish law and from astrology, seems to have spread very widely throughout the world. It certainly reached India by the fifth century, and apparently even went as far as Japan. References to a seven-day week have been found in Japanese literature of 1000 years ago. When referring to the Gregorian calendar, the Japanese also give the days of the week the names of the planets, in exactly the same order as in the French and Italian calendars, that is, Moon, Mars, Mercury, Jupiter, Venus, Saturn, and Sun.

Colson (1926) gives a thorough discussion of what he calls the "planetary week" and explains the more plausible of two analyses for the particular order derived from a history of Rome written by Dion Cassius in the early third century. The natural ordering of the nonfixed heavenly bodies, from a geocentric point of view, is determined by the rapidity with which they move among the stars. The Moon is by far the fastest, moving about 13° per day, whereas the Sun moves only 1° per day. Mercury and Venus sometimes loop around the Sun, and hence move faster than the Sun. When these bodies are arranged from slowest to fastest in their movement across the sky, as seen from the earth, the order is: Saturn, Jupiter, Mars, Sun, Venus, Mercury, Moon. Taking every third one in cyclic order, starting with the Sun, we get the arrangement Sun, Moon, Mars, Mercury, Jupiter, Venus, Saturn, which is the cyclic order of the days of the week. Dion Cassius explains this order as follows: The planets take turns keeping one-hour tours of guard duty, so to speak. In any such cyclic arrangement, the planet that was on duty during the fourth hour of each day will be on duty during the first hour of the next day. The days are named after the planet that is on watch at sunrise. Since there are 24 hours in the day and seven planets, you can see that number of the first hour of successive days will be 1, 25, 49, 73, 97, 121, and 145. Up to multiples of seven, these numbers are equal to 1, 4, 7, 3, 6, 2, 5 respectively, and that is the cyclic order of our weekdays.

In his *Aryabhatiya* the Hindu writer Aryabhata I (476–550) uses the planetary names for the days of the week and explains the correlation in a manner consistent with the hypothesis of Dion Cassius. He writes:

[C]ounting successively the fourth in the order of their swiftness they become the Lords of the days from sunrise. [Clark, 1930, p. 56]

The hypothesis of Dion Cassius is plausible and has been widely accepted for centuries. More than 600 years ago the poet Geoffrey Chaucer wrote a treatise on the astrolabe (Chaucer, 1391) in which he said

The firste houre inequal of every Saturday is to Saturne, and the seconde to Jupiter, the thirde to Mars, the fourthe to the sonne,

the fite to Venus, the sixte to Mercurius, the seventh to the mone.
And then ageyn the 8 houre is to Saturne, the 9 is to Jupiter, the 10
to Mars, the 11 to the sonne, the 12 to Venus. . . And in this manner
succedith planete under planete fro Saturne unto the mone, and
fro the mone up ageyn to Saturne. [Chaucer, 1391, Robinson, p.
553]

Chaucer made references to these hours in the *Canterbury Tales*.⁹ Thus, all this planetary lore and the seven-day week have their origin in the sexagesimal system of counting and the division of the day into 24 hours, which we know is characteristic of ancient Mesopotamia. But if the Mesopotamians had used a decimal system and divided their days into 10 hours, the days would still occur in the same order, since 10 and 24 are congruent modulo 7.

Questions and problems

5.1. Find an example, different from those given in the text, in which English grammar makes a distinction between a set of two and a set of more than two objects.

5.2. Consider the following three-column list of number names in English and Russian. The first column contains the cardinal numbers (those used for counting), the second column the ordinal numbers (those used for ordering), and the third the fractional parts. Study and compare the three columns. The ordinal numbers and fractions and the numbers 1 and 2 are grammatically adjectives in Russian. They are given in the feminine form, since the fractions are always given that way in Russian, the noun *dolya*, meaning *part* or *share*, always being understood. If you know another language, prepare a similar table for that language, then describe your observations and inferences. What does the table suggest about the origin of counting?

English			Russian		
one	first	whole	odna	pervaya	tselaya
two	second	half	dve	vtoraya	polovina
three	third	third	tri	tret'ya	tret'
four	fourth	fourth	chetyre	chetvyortaya	chetvert'
five	fifth	fifth	pyat'	pyataya	pyataya
six	sixth	sixth	shest'	shestaya	shestaya

5.3. How do you account for the fact that the ancient Greeks used a system of counting and calculating that mirrored the notation found in Egypt, whereas in their astronomical measurements they borrowed the sexigesimal system of Mesopotamia? Why were they apparently blind to the computational advantages of the place-value system used in Mesopotamia?

5.4. A tropical year is the time elapsed between successive south-to-north crossings of the celestial equator by the Sun. A sidereal year is the time elapsed between two successive conjunctions of the Sun with a given star; that is, it is the time required for the Sun to make a full circuit of the ecliptic path that it appears (from Earth) to follow among the stars each year. Because the celestial equator is rotating (one

⁹ See the 1928 edition edited by John Matthews Manly, published by Henry Holt and Company, New York, especially the third part of the Knight's Tale, pp. 198-213.

revolution in 26,000 years) in the direction opposite to the Sun's motion along the ecliptic, a tropical year is about 20 minutes shorter than a sidereal year. Would you expect the flooding of the Nile to be synchronous with the tropical year or with the sidereal year? If the flooding is correlated with the tropical year, how long would it take for the heliacal rising of Sirius to be one day out of synchronicity with the Nile flood? If the two were synchronous 4000 years ago, how far apart would they be now, and would the flood occur later or earlier than the heliacal rising of Sirius?

5.5. How many *Tzolkin* cycles are there in a Calendar Round?

5.6. The pattern of leap-year days in the Gregorian calendar has a 400-year cycle. Do the days of the week also recycle after 400 years?

5.7. (*The revised Julian calendar*) The Gregorian calendar bears the name of the Pope who decreed that it should be used. It was therefore adopted early in many countries with a Catholic government, somewhat later in Anglican and Protestant countries. Countries that are largely Orthodox in faith resisted this reform until the year 1923, when a council suggested that century years should be leap years only when they leave a remainder of 2 or 6 when divided by 9. (This reform was not mandated, but was offered as a suggestion, pending universal agreement among all Christians on a date for Easter.) This modification would retain only two-ninths of the century years as leap years, instead of one-fourth, as in the Gregorian calendar. What is the average number of days in a year of this calendar? How does it compare with the actual length of a year? Is it more or less accurate than the Gregorian calendar?

5.8. In constructing a calendar, we encounter the problem of measuring time. Measuring *space* is a comparatively straightforward task, based on the notion of congruent lengths. One can use a stick or a knotted rope stretched taut as a standard length and compare lengths or areas using it. Two lengths are congruent if each bears the same ratio to the standard length. In many cases one can move the objects around and bring them into coincidence. But what is meant by congruent *time intervals*? In what sense is the interval of time from 10:15 to 10:23 congruent to the time interval from 2:41 to 2:49?

5.9. It seems clear that the decimal place-value system of writing integers is *potentially infinite*; that is there is no limit on the size of number that can be written in this system. But in practical terms, there is always a largest number for which a name exists. In ordinary language, we can talk about trillions, quadrillions, quintillions, sextillions, septillions, octillions, and so on. But somewhere before the number 10^{60} is reached, most people (except Latin scholars) will run out of names. Some decades ago, a nephew of the American mathematician Edward Kasner (1878–1955) coined the name *googol* for the number 10^{100} , and later the name *googolplex* for $10^{10^{100}}$. This seems to be the largest number for which a name exists in English. Does there exist a positive integer for which no name *could* possibly be found, not merely an integer larger than all the integers that have been or will have been named before the human race becomes extinct? Give a logical argument in support of your answer. (And, while you are at it, consider what is meant by saying that an integer “exists.”)

CHAPTER 6

Calculation

In the present chapter we are going to look at processes that the modern calculator has rendered obsolescent, that is, the basic operations of arithmetic: addition, subtraction, multiplication, division, and the extraction of (square) roots. The word *obsolescent* is used instead of the more emphatic *obsolete* because these processes are still being taught to children in schools. But the skill that children are acquiring becomes weaker with every passing year. In fact, it has been at least 30 years since high-school students were actually taught to extract a square root. Even then, it was easier to consult a table of square roots than to carry out the error-prone, complicated operation of finding the root. Of course, what has caused these skills to fall out of use is the ready availability of hand-held calculators. This latest technological marvel is a direct continuation of earlier technology to ease the burden of concentration required in doing arithmetic, starting with counting rods and counting boards, then moving on through the abacus and the slide rule. The need to calculate has been a motivating force behind the development of mechanical methods of computation, and thus an important part of the history of mathematics. In this chapter we look first at the earliest methods developed for calculating, concentrating on multiplication and division (or, in the case of Egypt, processes equivalent to these) and the extraction of roots. We shall also look at three important motives for calculating: (1) commercial transactions involving labor, construction, and trade; (2) geometric problems of area and volume involving surveying and engineering; and (3) regulation of the calendar, especially finding important dates such as Easter.

1. Egypt

The richest source of information on Egyptian methods of calculation is the Ahmose (Rhind) Papyrus described in Chapter 2. After the descriptive title, the papyrus begins with the table of numbers shown in Fig. 1 below. In the modern world, we think of arithmetic as consisting of the four operations of addition, subtraction, multiplication, and division performed on whole numbers and fractions. We learn the rules for carrying out these operations in childhood and do them automatically, without attempting to prove that they are correct. The situation was different for the Egyptian. To the Egyptian, it seems, the fundamental operations were addition and *doubling*, and these operations were performed on whole numbers and *parts*. We need to discuss both the operations and the objects on which they were carried out.

Let us consider first the absence of multiplication and division as we know them. The tables you looked at in Problem 5.2 should have convinced you that there is something special about the number 2. We don't normally say "one-twoth"

for the result of dividing something in two parts. This linguistic peculiarity suggests that *doubling* is psychologically different from applying the general concept of multiplying in the special case when the multiplier is 2.

Next consider the absence of what we would call fractions. The closest Egyptian equivalent to a fraction is what we called a *part*. For example, what we refer to nowadays as the fraction $\frac{1}{7}$ would be referred to as “the seventh part.” This language conveys the image of a thing divided into seven equal parts arranged in a row and the seventh (and last) one being chosen. For that reason, according to van der Waerden (1963), there can be only *one* seventh part, namely the last one; there would be no way of expressing what we call the fraction $\frac{3}{7}$. An exception was the fraction that we call $\frac{2}{3}$, which occurs constantly in the Ahmose Papyrus. There was a special symbol meaning “the two parts” out of three. In general, however, the Egyptians used only *parts*, which in our way of thinking are *unit fractions*, that is, fractions whose numerator is 1. Our familiarity with fractions in general makes it difficult to see what the fuss is about when the author asks what must be added to the two parts and the fifteenth part in order to make a whole (Problem 21 of the papyrus). If this problem is stated in modern notation, it merely asks for the value of $1 - (\frac{1}{15} + \frac{2}{3})$, and of course, we get the answer immediately, expressing it as $\frac{4}{15}$. Both this process and the answer would have been foreign to the Egyptian, whose solution is described below.

To understand the Egyptians, we shall try to imitate their way of writing down a problem. On the other hand, we would be at a great disadvantage if our desire for authenticity led us to try to solve the entire problem using their notation. The best compromise seems to be to use our symbols for the whole numbers and express a *part* by the corresponding whole number with a bar over it. Thus, *the fifth part* will be written $\bar{5}$, *the thirteenth part* by $\bar{13}$, and so on. For “the two parts” ($\frac{2}{3}$) we shall use a double bar, that is, $\bar{\bar{3}}$.

1.1. Multiplication and division. Since the only operation other than addition and subtraction of integers (which are performed automatically without comment) is doubling, the problem that we would describe as “multiplying 11 by 19” would have been written out as follows:

19	1	*
38	2	*
76	4	
152	8	*
Result	209	11

Inspection of this process shows its justification. The rows are kept strictly in proportion by doubling each time. The final result can be stated by comparing the first and last rows: 19 is to 1 as 209 is to 11. The rows in the right-hand column that must be added in order to obtain 11 are marked with an asterisk, and the corresponding entries in the left-hand column are then added to obtain 209. In this way any two positive integers can easily be multiplied. The only problem that arises is to decide how many rows to write down and which rows to mark with an asterisk. But that problem is easily solved. You stop creating rows when the next entry in the right-hand column would be bigger than the number you are multiplying by (in this case 11). You then mark your last row with an asterisk, subtract the entry in its right-hand column (8) from 11 (getting a remainder of 3),

then move up and mark the next row whose right-hand column contains an entry not larger than this remainder (in this case the second row), subtract the entry in its right-hand column (2), from the previous remainder to get a smaller remainder (in this case 1), and so forth.

We shall refer to this general process of doubling and adding as *calculating*. What we call division is carried out in the same way, by reversing the roles of the two columns. For example, what we would call the problem of dividing 873 by 97 amounts to calculating with 97 so as to obtain 873. We can write it out as follows:

*	97	1
	194	2
	388	4
*	776	8
	873	9 Result.

The process, including the rules for creating the rows and deciding which ones to mark with an asterisk, is exactly the same as in the case of multiplication, except that now it is the left-hand column that is used rather than the right-hand column. We create rows until the next entry in the left-hand column would be larger than 873. We then mark the last row, subtract the entry in its left-hand column from 873 to obtain the remainder of 97, then look for the next row above whose left-hand entry contains a number not larger than 97, mark that row, and so on.

1.2. "Parts". Obviously, the second use of the two-column system can lead to complications. While in the first problem we can always express any positive integer as a sum of powers of 2, the second problem is a different matter. We were just lucky that we happened to find multiples of 97 that add up to 873. If we hadn't found them, we would have had to deal with those *parts* that have already been discussed. For example, if the problem were "calculate with 12 so as to obtain 28," it might have been handled as follows:

	12	1
*	24	2
	8	$\overline{3}$
*	4	$\overline{3}$
	28	$2\overline{3}$ Result.

What is happening in this computation is the following. We stop creating rows after 24 because the next entry in the left-hand column (48) would be bigger than 28. Subtracting 24 from 28, we find that we still need 4, yet no 4 is to be found. We therefore go back to the first row and multiply by $\frac{2}{3}$, getting the row containing 8 and $\overline{3}$. Dividing by 2 again gets a 4 in the left-hand column. We then have the numbers we need to get 28, and the answer is expressed as $2\overline{3}$. Quite often the first multiplication by a *part* involves the two-thirds part $\overline{3}$. The scribes probably began with this part instead of one-half for the same reason that a carpenter uses a plane before sandpaper: the work goes faster if you take bigger "bites."

The parts that are negative powers of 2 play a special role. When applied to a hekat of grain, they are referred to as the *Horus-eye* parts.¹ Since $1/2 + 1/4 +$

¹ According to Egyptian legend, the god Horus lost an eye in a fight with his uncle, and the eye was restored by the god Thoth. Each of these fractions was associated with a particular part of Horus' eye.

$1/8 + 1/16 + 1/32 + 1/64 = 63/64$, the scribes apparently saw that unity could be restored (approximately), as Horus' eye was restored, by using these parts. The fact that (in our terms) 63 occurs as a numerator, shows that division by 3, 7, and 9 is facilitated by the use of the Horus-eye series. In particular, since $1/7 = (1/7) \cdot ((63/64) + 1/64) = 9/64 + 1/448 = 8/64 + 1/64 + 1/448$, the seventh part could have been written as $\overline{8} \overline{64} \overline{448}$. In this way, the awkward seventh part gets replaced by the better-behaved Horus-eye fractions, plus a corrective term (in this case $\overline{448}$, which might well be negligible in practice. Five such replacements are implied, though not given in detail, in the Akhmim Wooden Tablet.² As another example, since $64 = 4 \cdot 13 + 8 + 4$, we find that $\overline{13}$ becomes $\overline{16} \overline{104} \overline{208}$.

There are two more complications that arise in doing arithmetic the Egyptian way. The first complication is obvious. Since the procedure is based on doubling, but the double of a *part* may not be expressible as a part, how does one "calculate" with parts? It is easy to double, say, the twenty-sixth part: The double of the twenty-sixth part is the thirteenth part. If we try to double again, however, we are faced with the problem of doubling a part involving an odd number. The table at the beginning of the papyrus gives the answer: The double of the thirteenth part is the eighth part plus the fifty-second part plus the one hundred fourth part. In our terms this tabular entry expresses the fact that

$$\frac{2}{13} = \frac{1}{8} + \frac{1}{52} + \frac{1}{104}.$$

Gillings (1972, p. 49) lists five precepts apparently followed by the compiler of this table in order to make it maximally efficient for use. The most important of these are the following three. One would like each double (1) to have as few terms as possible, (2) with each term as small as possible (that is, the "denominators" as small as possible), and (3) with even "denominators" rather than odd ones. These principles have to be balanced against one another, and the table in Fig. 1 represents the resulting compromise. However, Gillings' principles are purely negative ones, telling what *not* to do. The positive side of creating such a table is to find simple patterns in the numbers. One pattern that occurs frequently is illustrated by the double of $\overline{5}$, and amounts to the identity $2/p = 1/((p+1)/2) + 1/(p(p+1)/2)$. Another, illustrated by the double of $\overline{13}$, probably arises from the Horus-eye representation of the original part.

With this table, which gives the doubles of all parts involving an odd number up to 101, calculations involving parts become feasible. There remains, however, one final complication before one can set out to solve problems. The calculation process described above requires subtraction at each stage in order to find what is lacking in a given column. When the column already contains *parts*, this leads to the second complication: the problem of *subtracting parts*. (*Adding parts* is no problem. The author merely writes them one after another. The sum is condensed if, for example, the author knows that the sum of $\overline{3}$ and $\overline{6}$ is $\overline{2}$.) This technique, which is harder than the simple procedures discussed above, is explained in the papyrus itself in Problems 21 to 23. As mentioned above, Problem 21 asks for the parts that must be added to the sum of $\overline{3}$ and $\overline{15}$ to obtain 1. The procedure used to solve this problem is as follows. Begin with the two parts in the first row:

² See <http://www.mathworld.com/AkhmimWoodenTablet.html>. In a post to the history of mathematics mailing list in December 2004 the author of that article, Milo Gardner, noted that recent analysis of this tablet has upset a long-held belief about the meaning of a certain term in these equations.

5	3 15	55	30 318 795
7	4 28	57	38 114
9	6 18	59	36 236 531
11	6 66	61	40 244 488 610
13	8 52 104	63	42 126
15	10 30	65	39 195
17	12 51 68	67	40 335 536
19	12 76 114	69	46 138
21	14 42	71	40 568 710
23	12 276	73	60 219 292 365
25	15 75	75	50 150
27	18 54	77	44 308
29	24 58 174 232	79	60 237 316 790
31	20 124 155	81	54 162
33	22 66	83	60 332 415 498
35	30 42	85	51 255
37	24 111 296	87	58 174
39	26 78	89	60 356 534 890
41	24 246 328	91	70 130
43	42 86 129 301	93	62 186
45	30 90	95	60 380 570
47	30 141 470	97	56 679 776
49	28 196	99	66 198
51	34 102	101	101 202 303 606

FIGURE 1. Doubles of unit fractions in the Ahmose Papyrus

$$\bar{3} \quad \bar{15} \quad 1.$$

Now the problem is to see what must be added to the two terms on the left-hand side in order to obtain the right-hand side. Preserving proportions, the author multiplies the row by 15, getting

$$10 \quad 1 \quad 15$$

It is now clear that when the problem is "magnified" by a factor of 15, we need to add 4 units. Therefore, the only remaining problem is, as we would put it, to divide 4 by 15, or in language that may reflect better the thought process of the author, to "calculate with 15 so as to obtain 4." This operation is carried out in the usual way:

$$\begin{array}{rcl}
 15 & 1 & \\
 1 & \bar{15} & \\
 2 & \bar{10} \bar{30} & \text{[from the table]} \\
 4 & \bar{5} \bar{15} & \text{Result.}
 \end{array}$$

Thus, the parts that must be added to the sum of $\bar{3}$ and $\bar{15}$ in order to reach 1 are $\bar{5}$ and $\bar{15}$. This "subroutine," which is essential to make the system of computation work, was written in red ink in the manuscripts, as if the writers distinguished between computations made within the problem to find the answer and computations made in order to operate the system. Having learned how to complement

(subtract) parts, what are called *hau* (or *aha*) computations by the author, one can confidently attack any arithmetic problem whatsoever. Although there is no single way of doing these problems, specialists in this area have detected systematic procedures by which the table of doubles was generated and patterns in the solution of problems that indicate, if not an algorithmic procedure, at least a certain habitual approach to such problems.

Let us now consider how these principles are used to solve a problem from the papyrus. The one we pick is Problem 35, which, translated literally and misleadingly, reads as follows:

Go down I times 3. My third part is added to me. It is filled. What is the quantity saying this? ·

To clarify: This problem asks for a number that yields 1 when it is tripled and the result is then increased by the third part of the original number. In other words, "calculate with $3\bar{3}$ so as to obtain 1." The solution is as follows:

$$\begin{array}{rcl}
 3\bar{3} & 1 & \\
 10 & 3 & \text{[multiplied by 3]} \\
 5 & 1\bar{2} & \\
 1 & 5\bar{10} & \text{Result.}
 \end{array}$$

1.3. Practical problems. One obvious application of calculation in everyday life is in surveying, where one needs some numerical way of comparing the sizes of areas of different shapes. This application is discussed in Chapter 9. The papyrus also contains several problems that involve proportion in the guise of the slope of pyramids and the strength of beer. Both of these concepts involve what we think of as a ratio, and the technique of finding the fourth element in a proportion by the procedure once commonly taught to grade-school students and known as the *Rule of Three*. It is best explained by a sample question. If three bananas cost 69 cents, what is the cost of five bananas? Here we have three numbers: 3, 69, and 5. We need a fourth number that has the same ratio to 69 that 5 has to 3, or, equivalently, the same ratio to 5 that 69 has to 3. The rule says that such a number is $69 \times 5 \div 3 = 105$. Since the Egyptian procedure for multiplication was based on an implicit notion of proportion, such problems yield easily to the Egyptian techniques. We shall reserve the discussion of pyramid slope problems until we examine Egyptian geometry in Chapter 9. Several units of weight are mentioned in these problems, but the measurement we shall pay particular attention to is a measure of the dilution of bread or beer. It is called a *pesu* and defined as the number of loaves of bread or jugs of beer obtained from one *hekat* of grain. A hekat was slightly larger than a gallon, 4.8 liters to be precise. Just how much beer or bread it would produce under various circumstances is a technical matter that need not concern us. The thing we need to remember is that the number of loaves of bread or jugs of beer produced by a given amount of grain equals the *pesu* times the number of hekats of grain. A large *pesu* indicates weak beer or bread. In the problems in the Ahmose Papyrus the *pesu* of beer varies from 1 to 4, while that for bread varies from 5 to 45.

Problem 71 tells of a jug of beer produced from half a hekat of grain (thus its *pesu* was 2). One-fourth of the beer is poured off and the jug is topped up with water. The problem asks for the new *pesu*. The author reasons that the eighth part of a hekat of grain was removed, leaving (in his terms) $\bar{4}\bar{8}$, that is, what we would



FIGURE 2. The Shang numerals.

call $\frac{3}{8}$ of a hekat of grain. Since this amount of grain goes into one jug, it follows that the *pesu* of that beer is what we call the *reciprocal* of that number, namely $2\frac{2}{3}$. The author gives this result immediately, apparently assuming that by now the reader will know how to “calculate with $4\frac{8}{8}$ until 1 is reached.” The Rule of Three procedure is invoked in Problem 73, which asks how many loaves of 15-*pesu* bread are required to provide the same amount of grain as 100 loaves of 10-*pesu* bread. The answer is found by dividing 100 by 10, then multiplying by 15, which is precisely the Rule of Three.

2. China

In contrast to the Egyptians, who computed with ink on papyrus, the ancient Chinese, starting in the time of the Shang dynasty, used rods representing numerals to carry out computations. Chinese documents from the second century BCE mention the use of counting rods, and a set of such rods from the first century BCE was discovered in 1970. The rods can be arranged to form the Shang numerals (Fig. 2) and thereby represent decimal digits. They were used in conjunction with a counting board, which is a board ruled into squares so that each column (or row, depending on the direction of writing) represents a particular item. In pure computations, the successive rows in the board indexed powers of 10. These rods could be stacked to represent any digit from 1 to 9. Since they were placed on a board in rows and columns, the empty places are logically equivalent to a use of 0, but not psychologically equivalent. The use of a circle for zero in China is not found before the thirteenth century. On the other hand, according to Lam and Ang (1987, p. 102), the concept of negative numbers (*fu*), represented by black rods instead of the usual red ones for positive numbers (*cheng*), was also present as early as the fourth century BCE.

It is difficult to distinguish between, say, 22 (|||) and 4 (||||) if the rods are placed too close together. To avoid that difficulty, the Chinese rotated the rods in alternate rows through a right angle, in effect using a positional system based on 100 rather than 10. Since this book is being published in a language that is read from left to right, then from top to bottom, we shall alternate columns rather than rows. In our exposition of the system the number 22 becomes ==|| and 4 remains |||. The Shang numerals are shown in Fig. 2, the top row being used to represent digits multiplied by an even power of 10 and the bottom row digits multiplied by an odd power of 10.

Addition and subtraction with rods representing Shang numerals are obvious operations. Multiplication and division require somewhat more work, and those procedures are explained in the *Sun Zi Suan Jing*.

Except that multiplication was carried out starting with the largest denominations rather than the smallest, the procedure for multiplying digits and carrying resembles all other systems for multiplying. Using numerals in place of the rods,

we can illustrate the multiplication $324 \cdot 29$ as follows:

$$\begin{array}{ccccccc}
 \begin{array}{r} 3 \ 2 \ 4 \\ \rightarrow 6 \\ 2 \ 9 \end{array} &
 \begin{array}{r} 3 \ 2 \ 4 \\ \rightarrow 8 \ 7 \\ 2 \ 9 \end{array} &
 \begin{array}{r} 3 \ 2 \ 4 \\ \rightarrow 8 \ 7 \\ 2 \ 9 \end{array} &
 \begin{array}{r} 2 \ 4 \\ \rightarrow 9 \ 1 \\ 2 \ 9 \end{array} &
 \begin{array}{r} 2 \ 4 \\ \rightarrow 9 \ 1 \\ 2 \ 9 \end{array} \\
 \\
 \begin{array}{r} 2 \ 4 \\ \rightarrow 9 \ 2 \ 8 \\ 2 \ 9 \end{array} &
 \begin{array}{r} 4 \\ \rightarrow 9 \ 2 \ 8 \\ 2 \ 9 \end{array} &
 \begin{array}{r} 4 \\ \rightarrow 9 \ 3 \ 6 \\ 2 \ 9 \end{array} &
 \begin{array}{r} 4 \\ \rightarrow 9 \ 3 \ 9 \ 6 \\ 2 \ 9 \end{array} &
 \begin{array}{r} 4 \\ \rightarrow 9 \ 3 \ 9 \ 6 \\ 2 \ 9 \end{array}
 \end{array}$$

The sequence of operations is very easy to understand from this illustration. First, the larger number is written on top (on the right in Chinese, of course, since the writing is vertical). The smaller number is written on the bottom (actually on the left), with its units digit opposite the largest digit of the larger number. Then, working always from larger denominations to smaller, we multiply the digits one at a time and enter the products between the two numbers. Once a digit of the larger number has been multiplied by all the digits of the smaller one, it is "erased" (the rods are picked up), and the rods representing the smaller number are moved one place to the right (actually downward). At that point, the process repeats until all the digits have been multiplied. When that happens, the last digit of the larger number and all the digits of the smaller number are picked up, leaving only the product.

Long division was carried out in a similar way. The partial quotients were kept in the top row, and the remainder at each stage occupied the center row (with the same caveat as above, that rows are actually columns in Chinese writing). For example, to get the quotient $438 \div 7$, one proceeds as follows.

$$\begin{array}{ccccccc}
 & & & 6 & & 6 & & 6 \ 2 \\
 4 \ 3 \ 8 & \rightarrow & 4 \ 3 \ 8 & \rightarrow & 4 \ 3 \ 8 & \rightarrow & 1 \ 8 & \rightarrow & 1 \ 8 & \rightarrow & 4 \\
 & & 7 & & 7 & & 7 & & 7 & & 7
 \end{array}$$

The first step here is merely a statement of the problem. The procedure begins with the second step, where the divisor (7) is moved to the extreme left, then moved rightward until a division is possible. Thereafter, one does simple divisions, replacing the dividend by the remainder at each stage. The original dividend can be thought of as the remainder of a fictitious "zeroth" division. Except for the "erasures" when the rods are picked up, the process looks very much like the algorithm taught to school children in the United States. The final display allows the answer to be read off: $438 \div 7 = 62\frac{4}{7}$. It would be only a short step to replace this last common fraction by a decimal; all one would have to do is continue the algorithm as if there were zeros on the right of the dividend. However, no such procedure is described in the *Sun Zi Suan Jing*. Instead, the answer is expressed as an integer plus a proper fraction.

2.1. Fractions and roots. The early Chinese way of handling fractions is much closer to our own ideas than that of the Egyptians. The *Sun Zi Suan Jing* gives a procedure for reducing fractions that is equivalent to the familiar Euclidean algorithm for finding the greatest common divisor of two integers. The rule is to subtract the smaller number from the larger until the difference is smaller than the originally smaller number. Then begin subtracting the difference from the smaller number. Continue this procedure until two equal numbers are obtained. That number can then be divided out of both numerator and denominator.

With this procedure for reducing fractions to lowest terms, a complete and simple theory of computation with fractions is feasible. Such a theory is given in the *Sun Zi Suan Jing*, including the standard procedure for converting a mixed number to an improper fraction and the procedures for adding, subtracting, multiplying, and dividing fractions. Thus, the Chinese had complete control over the system of rational numbers, including, as we shall see below, the negative rational numbers.

At an early date the Chinese dealt with roots of integers, numbers like $\sqrt{355}$, which we now know to be irrational; and they found mixed numbers as approximations when the integer is not a perfect square. In the case of $\sqrt{355}$, the approximation would have been given as $18\frac{31}{36}$. (The denominator is always twice the integer part, as a result of the particular algorithm used.)

From arithmetic to algebra. Sooner or later, constantly solving problems of more and more complexity in order to find unknown quantities leads to the systematization of ways of imagining operations performed on a “generic” number (unknown). When the point arises at which a name or a symbol for an unknown number is invented, so that *expressions* can be written representing the result of operations on the unknown number, we may take it that algebra has arisen. There is a kind of twilight zone between arithmetic and algebra, in which certain problems are solved imaginatively without using symbols for unknowns, but later are seen to be easily solvable by the systematic methods of algebra. A good example is Problem 15 of Chapter 3 of the *Sun Zi Suan Jing*, which asks how many carts and how many people are involved, given that there are two empty carts (and all the others are full) when people are assigned three to a cart, but nine people have to walk if only two are placed in each cart. We would naturally make this a problem in two linear equations in two unknowns: If x is the number of people and y the number of carts, then

$$\begin{aligned}x &= 3(y - 2), \\x &= 2y + 9.\end{aligned}$$

However, that would be using algebra, and Sun Zi does not quite do that in this case. His solution is as follows:

Put down 2 carts, multiply by 3 to give 6, add 9, which is the number of persons who have to walk, to obtain 15 carts. To find the number of persons, multiply the number of carts by 2 and add 9, which is the number of persons who have to walk.

Probably the reasoning in the first sentence here is pictorial. Imagine each cart filled with three people. When loaded in this way, the carts would accommodate all the “real” people in the problem, plus six “fictitious” people, since we are given that two carts would be empty if the others each carried three people. Let us imagine then, that six of the carts contain two real people and one fictitious person, while the others contain three real people. Now imagine one person removed from each cart, preferably a fictitious person if possible. The number of people removed would obviously be equal to the number of carts. The six fictitious people would then be removed, along with the nine real people who have to walk when there are only two people in each cart. It follows that there must be 15 carts. Finding the number of people is straightforward once the number of carts is known.

2.2. The *Jiu Zhang Suanshu*. This work is the most fundamental of the early Chinese mathematical classics. For the most part, it assumes that the methods of calculation explained in the *Sun Zi Suan Jing* are known and applies them to problems very similar to those discussed in the Ahmose Papyrus. In fact, Problems 5, 7, 10, and 15 from the first chapter reappear as the first four problems of Chapter 2 of the *Sun Zi Suan Jing*. As its title implies, the book is divided into nine chapters. These nine chapters contain a total of 246 problems. The first eight of these chapters discuss calculation and problems that we would now solve using linear algebra. The last chapter is a study of right triangles. The first chapter, whose title is "Rectangular Fields," discusses how to express the areas of fields given their sides. Problem 1, for example, asks for the area of a rectangular field that is 15 *bu* by 16 *bu*.³ The answer, we see immediately, is 240 "square *bu*." However, the Chinese original does not distinguish between linear and square units. The answer is given as "1 *mu*." The *Sun Zi Suan Jing* explains that as a unit of *length*, 1 *mu* equals 240 *bu*. This ambiguity is puzzling, since a *mu* is both a length equal to 240 *bu* and the area of a rectangle whose dimensions are 1 *bu* by 240 *bu*. It would seem more natural for us if 1 *mu* of area were represented by a square of side 1 *mu*. If these units were described consistently, a square of side 1 linear *mu* would have an area equal to 240 "areal" *mu*. That there really is such a consistency appears in Problems 3 and 4, in which the sides are given in *li*. Since 1 *li* equals 300 *bu* (that is, 1.25 *mu*), to convert the area into *mu* one must multiply the lengths of the sides in *li*, then multiply by $1.25^2 \cdot 240 = 375$. In fact, the instructions say to multiply by precisely that number.

Rule of Three problems. Chapter 2 ("Millet and Rice") of the *Jiu Zhang Suanshu* contains problems very similar to the *pesu* problems from the Ahmose Papyrus. The proportions of millet and various kinds of rice and other grains are given as empirical data at the beginning of the chapter. If the Ahmose Papyrus were similarly organized into chapters, the chapter in it corresponding to this chapter would be called "Grain and Bread." Problems of the sort studied in this chapter occur frequently in all commercial transactions in all times. In the United States, for example, a concept analogous to *pesu* is the *unit price* (the number of dollars the merchant will obtain by selling 1 unit of the commodity in question). This number is frequently printed on the shelves of grocery stores to enable shoppers to compare the relative cost of purchasing different brands. Thus, the practicality of this kind of calculation is obvious. The 46 problems in Chapter 2, and also the 20 problems in Chapter 3 ("Proportional Distribution") of the *Jiu Zhang Suanshu* are of this type, including some extensions of the Rule of Three. For example, Problem 20 of Chapter 3 asks for the interest due on a loan of 750 *qian* repaid after 9 days if a loan of 1000 *qian* earns 30 *qian* interest each month (a month being 30 days). The result is obtained by forming the product 750 *qian* times 30 *qian* times 9 days, then dividing by the product 1000 *qian* times 30 days, yielding $6\frac{3}{4}$ *qian*. Here the product of the monthly interest on a loan of 1 *qian* and the number of days the loan is outstanding, divided by 30, forms the analog of the *pesu* for the loan, that is, the number of *qian* of interest produced by each *qian* loaned. Further illustrations are given in the problems at the end of the chapter.

³ One *bu* is 600,000 *hu*, a *hu* being the diameter of a silk thread as it emerges from a silkworm. Estimates are that 1 *bu* is a little over 2 meters.

Chapter 6 ("Fair Transportation") is concerned with the very important problem of fair allocation of the burdens of citizenship. The Chinese idea of fairness, like that in many other places, including modern America, involves direct proportion. For example, Problem 1 considers a case of collecting taxes in a given location from four counties lying at different distances from the collection center and having different numbers of households. To solve this problem a constant of proportionality is assigned to each county equal to the number of its households divided by its distance from the collection center. The amount of tax (in millet) each county is to provide is its constant divided by the sum of all the constants of proportionality and multiplied by the total amount of tax to be collected. The number of carts (of a total prescribed number) to be provided by each county is determined the same way. The data in the problem are as follows.

County	Number of Households	Distance to Collection Center
A	10,000	8 days
B	9,500	10 days
C	12,350	13 days
D	12,200	20 days

A total of 250,000 *hu* of millet were to be collected as tax, using 10,000 carts. The proportional parts for the four counties were therefore 1250, 950, 950, and 610, which the author reduced to 125, 95, 95, and 61. These numbers total 376. It therefore followed that county A should provide $\frac{125}{376} \cdot 250,000$ *hu*, that is, approximately 83,111.7 *hu* of millet and $\frac{125}{376} \cdot 10,000$, or 3324 carts. The author rounded off the tax to three significant digits, giving it as 83,100 *hu*.

Along with these administrative problems, the 28 problems of Chapter 6 also contain some problems that have acquired an established place in algebra texts throughout the world and will be continue to be worked by students as long as there are teachers to require it. For example, Problem 26 considers a pond used for irrigation and fed by pipes from five different sources. Given that these five canals, each "working" alone, can fill the pond in $\frac{1}{3}$, 1, $2\frac{1}{2}$, 3, and 5 days, the problem asks how long all five "working" together will require to fill it. The author realized that the secret is to add the rates at which the pipes "work" (the reciprocals of the times they require individually to fill the pond), then take the reciprocal of this sum, and this instruction is given. The answer is $1/(3 + 1 + 2/5 + 1/3 + 1/5) = 15/74$.

3. India

We have noted a resemblance between the mathematics developed in ancient Egypt and that developed in ancient China. We should not be surprised at this resemblance, since these techniques arose in response to universal needs in commerce, industry, government, and society. They form a universal foundation for mathematics that remained at the core of any practical education until very recent times. Only the widespread use of computers and computer graphics has, over the past two decades, made these skills obsolete, just as word processors have made it unimportant to develop elegant handwriting.

To avoid repetition, we simply note that much of the Hindu method of computation is similar to what is now done or what is discussed in other sections of this chapter. A few unusual aspects can be noted, however. Brahmagupta gives the standard rules for handling common fractions. However, his arithmetic contains some original ways of looking at many things that we take for granted. For

example, to do a long division with remainder, say, $\frac{750}{22}$, he would look for the next number after 22 that divides 750 evenly (25) and write

$$\frac{750}{22} = \frac{750}{25} + \left(\frac{750}{25}\right)\frac{3}{22},$$

that is,

$$\frac{750}{22} = 30\left(1 + \frac{3}{22}\right) = 30 + \frac{90}{22} = 34\frac{1}{11}.$$

Beyond these simple operations, he also codifies the methods of taking square and cube roots, and he states clearly the Rule of Three (Colebrooke, 1817, p. 283). Brahmagupta names the three terms the “argument,” the “fruit,” and the “requisition,” and points out that the argument and the requisition must be the same kind of thing. The unknown number he calls the “produce,” and he gives the rule that the produce is the requisition multiplied by the fruit and divided by the argument.

4. Mesopotamia

Cuneiform tablets from the site of Senkerch (also known as Larsa), kept in the British Museum, contain tables of products, reciprocals, squares, cubes, square roots, and cube roots of integers. It appears that the people who worked with mathematics in Mesopotamia learned by heart, just as we do, the products of all the small integers. Of course, for them a theoretical multiplication table would have to go as far as 59×59 , and the consequent strain on memory would be large. That fact may account for the existence of so many written tables. Just as most of us learn, without being required to do so, that $\frac{1}{3} = 0.3333\dots$, the Mesopotamians wrote their fractions as sexagesimal fractions and came to recognize certain reciprocals, for example $\frac{1}{9} = 0;6,40$. With a system based on 30 or 60, all numbers less than 10 except 7 have terminating reciprocals. In order to get a terminating reciprocal for 7 one would have to go to a system based on 210, which would be far too complicated.

Even with base 60, multiplication can be quite cumbersome, and historians have conjectured that calculating devices such as an abacus might have been used, although none have been found. Høyrup (2002) has analyzed the situation by considering the errors in two problems on Old Babylonian cuneiform tablets and deduced that any such device would have had to be some kind of counting board, in which terms that were added could not be identified and subtracted again (like pebbles added to a pile).

Not only are sexagesimal fractions handled easily in all the tablets, the concept of a square root occurs explicitly, and actual square roots are approximated by sexagesimal fractions, showing that the mathematicians of the time realized that they hadn't been able to make these square roots come out even. Whether they realized that the square root would never come out even is not clear. For example, text AO 6484 (the AO stands for *Antiquités Orientales*) from the Louvre in Paris contains the following problem on lines 19 and 20:

The diagonal of a square is 10 Ells. How long is the side? [To find the answer] multiply 10 by 0;42,30. [The result is] 7;5.

Now $0;42,30$ is $\frac{42}{60} + \frac{30}{3600} = \frac{17}{24} = 0.7083$, approximately. This is a very good approximation to $1/\sqrt{2} \approx 0.7071$, and the answer $7;5$ is, of course, $7\frac{1}{12} = 7.083 = 10 \cdot 0.7083$. It seems that the writer of this tablet knew that the ratio of the side of a square to its diagonal is approximately $\frac{17}{24}$. The approximation to $\sqrt{2}$ that arises from what is now called the *Newton-Raphson* method, starting from $\frac{3}{2}$ as the

first approximation, turns up the number $\frac{17}{12}$ as the next approximation. Thus the fraction $\frac{17}{24}$ represents an approximation to $\frac{\sqrt{2}}{2} = \frac{1}{\sqrt{2}}$. The method of approximating square roots can be understood as an averaging procedure. In the present case, it works as follows. Since 1 is smaller than $\sqrt{2}$ and 2 is larger, let their average be the first approximation, that is, $\frac{3}{2}$. This number happens to be larger than $\sqrt{2}$, but it is not necessary to know that fact to improve the approximation. Whether it errs by being too large or too small, the result of dividing 2 by this number will err in the other direction. Thus, since $\frac{3}{2}$ is too large to be $\sqrt{2}$, the quotient $\frac{2}{3/2} = \frac{4}{3}$ is too small. The average of these two numbers will be closer to $\sqrt{2}$ than either number;⁴ the second approximation to $\sqrt{2}$ is then $\frac{1}{2}(\frac{3}{2} + \frac{4}{3}) = \frac{17}{12}$. Again, whether this number is too large or too small, the number $\frac{2}{17/12} = \frac{24}{17}$ will err in the opposite direction, so that we can average the two numbers again and continue this process as long as we like. Of course, we cannot *know* that this procedure was used to get the approximate square root unless we find a tablet that says so.

The writers of these tablets realized that when numbers are combined by arithmetic operations, it may be of interest to know how to recover the original data from the result. This realization is the first step toward attacking the problem of inverting binary operations. Although we now solve such problems by solving quadratic equations, the Mesopotamian approach was more like the Chinese approach described above. That is, certain arithmetic processes that could be pictured were carried out, but what we call an equation was not written explicitly. With every pair of numbers, say 13 and 27, they associated two other numbers: their average $(13 + 27)/2 = 20$ and their *semidifference*⁵ $(27 - 13)/2 = 7$. The average and semidifference can be calculated from the two numbers, and the original data can be calculated from the average and semidifference. The larger number (27) is the sum of the average and semidifference: $20 + 7 = 27$, and the smaller number (13) is their difference: $20 - 7 = 13$. The realization of this mutual connection makes it possible essentially to "change coordinates" from the number pair (a, b) to the pair $((a + b)/2, (a - b)/2)$.

At some point lost to history some Mesopotamian mathematician came to realize that the product of two numbers is the difference of the squares of the average and semidifference: $27 \cdot 13 = (20)^2 - 7^2 = 351$ (or 5, 51 in Mesopotamian notation). This principle made it possible to recover two numbers knowing their sum and product or knowing their difference and product. For example, given that the sum is 10 and the product is 21, we know that the average is 5 (half of the sum), hence that the square of the semidifference is $5^2 - 21 = 4$. Therefore, the semidifference is 2, and the two numbers are $5 + 2 = 7$ and $5 - 2 = 3$. Similarly, knowing that the difference is 9 and the product is 52, we conclude that the semidifference is 4.5 and the square of the average is $52 + (4.5)^2 = 72.25$. Hence the average is $\sqrt{72.25} = 8.5$.

⁴ The error made by the average is half of the *difference* of the two errors.

⁵ This word is coined because English contains no one-word description of this concept, which must otherwise be described as half of the difference of the two numbers. It is clear from the way in which the semidifference occurs constantly that the writers of these tablets automatically looked at this number along with the average when given two numbers as data. However, there seems to be no word in the Akkadian, Sumerian, and ideogram glossary given by Neugebauer to indicate that the writers of the clay tablets had a special word for these concepts. But at the very least, they were trained to calculate these numbers when dealing with this type of problem. In the translations given by Neugebauer the average and semidifference are obtained one step at a time, by first adding or subtracting the two numbers, then taking half of the result.

Therefore, the two numbers are $8.5 + 4.5 = 13$ and $8.5 - 4.5 = 4$. The two techniques just illustrated occur constantly in the cuneiform texts, and seem to be procedures familiar to everyone, requiring no explanation.

5. The ancient Greeks

It is fairly obvious how to do addition and subtraction within the Greek system of numbering. Doing multiplication in modern notation, as you know, involves memorizing all products up to $9 \cdot 9$, and learning how to keep columns straight, plus carrying digits where necessary. With our place-value notation, we have little difficulty multiplying, say $23 \cdot 42$. For the ancient Greeks the corresponding problem of multiplying $\kappa'\gamma'$ by $\mu'\beta'$ was more complicated. It would be necessary to find four products: $\kappa' \cdot \mu'$, $\kappa' \cdot \beta'$, $\gamma' \cdot \mu'$, and $\gamma' \cdot \beta'$. The first of three of these require the one doing the calculation to keep in mind that κ' is 10 times β' and μ' is 10 times δ' .

These operations are easier for us, since we use the same nine digits in different contexts, keeping track of the numbers they represent by keeping the columns straight while multiplying. They were more difficult for the ancient Greeks, since going from 30 to 3 was not merely a matter of ignoring a zero; it involved a shift forward by 10 letters in the alphabet, from λ' to γ' . In addition to the carrying that we must do, the Greek calculator had to commit to memory 20 such alphabet shifts (10 by 10 letters, and 10 shifts by 20 letters), and, while computing, remember how many such shifts of each kind were performed, so as to know how many factors of 10 were being "stored" during the calculation.

The procedure is explained in the *Synagōgē* of the third-century mathematician Pappus of Alexandria. The surviving portion of this work begins in the middle of Book 2, explaining how to multiply quickly numbers that are multiples of 10. Pappus illustrates the procedure by the following example, in which number *names* are translated into English but number *symbols* are transcribed directly from the original. (See Fig. 3 in Chapter 5 for the numerical values represented by these Greek letters.)

Let the numbers be ν' , ν' , ν' , μ' , μ' , and λ' . Then the basic numbers will be ε' , ε' , ε' , δ' , δ' , and γ' . Their product will be ς' . Since there are ς' tens, and since ς' divided by four leaves a remainder of two, the product [of the six reduced numbers] will contain altogether one hundred myriads...and these ρ' myriads multiplied by the ς' units will make ξ' twofold [that is, "square"] myriads.

When we translate the problem into our notation, it becomes trivial. We are trying to find the product $50 \cdot 50 \cdot 50 \cdot 40 \cdot 40 \cdot 30$. We have no trouble factoring out the six zeros and rewriting the problem as $5 \cdot 5 \cdot 5 \cdot 4 \cdot 4 \cdot 3 \cdot 1,000,000 = 6000 \cdot 10^6 = 60 \cdot (10,000)^2$. Converting from 40 to 4 is considerably easier than converting from μ' to δ' . Pappus divided the number of tens by 4, since he was counting in myriads (10^4), and he expressed the answer as "*myriadōn ξ' diplōn*," that is, "of myriads 60 twofold," or, in better English, "60 twofold myriads." To describe what we now call the square of a number, the ancient Greeks had to extend the normal meaning of the word *diploos* (double). A nonmathematical reader of Pappus' Greek might

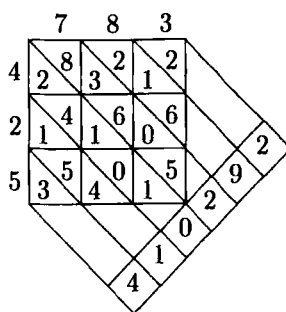


FIGURE 3. Multiplication with Hindu-Arabic numerals.

be inclined to think that the phrase *myrias diploos* meant $2 \cdot 10,000$ rather than $10,000^2$.

As you can see from this example, calculating (*logistikē*) was not quite so trivial for the Greeks as for us.

6. The Islamic world

It is well known that the numerals used all over the world today are an inheritance from both the Hindu and Arabic mathematicians of 1000 years ago. The Hindu idea of using nine symbols in a place-value system was known in what is now Iraq in the late seventh century, before that area became part of the Muslim Empire. In the late eighth century a scholar from India came to the court of Caliph al-Mansur with a work on Hindu astronomy using these numerals, and this work was translated into Arabic. An Arabic treatise on these numbers, containing the first known discussion of decimal fractions, was written by al-Uqlidisi (ca. 920–ca. 980).

Having inherited works from the time of Mesopotamia and also Greek and Hindu works that used the sexagesimal system in astronomy, the Muslim mathematicians of a thousand years ago also used that system. The sexagesimal system did not yield immediately to its decimal rival, and the technique of place-value computation developed in parallel in the two systems. Ifrah (2000, pp. 539–555) gives a detailed description of the long resistance to the new system. The sexagesimal system is mentioned in Arabic works of Abu'l-Wafa (940–988) and Kushar ben Laban (ca. 971–1029). It continued to appear in Arabic texts through the time of al-Kashi (1427), although the decimal system also occurs in the work of al-Kashi. In addition to the sexagesimal and decimal systems, the Muslim mathematicians used an elaborate system of finger reckoning. Some implementations of the decimal system require crossing out or erasing in the process of computation, and that was considered a disadvantage. Nevertheless, the superiority of decimal notation in computation was recognized early. For example, al-Daffa (1973, pp. 56–57) mentions that there are manuscripts still extant dating to the twelfth century, in which multiplication is performed by the very efficient method illustrated in Fig. 3 for the multiplication $524 \cdot 783 = 410,292$.

7. Europe

The system of Roman numerals that now remains in countries settled by Europeans is confined to a few cases where numbers have only to be read, not computed with.

For computations these cumbersome numerals were supplanted centuries ago by the Hindu-Arabic place-value decimal system. Before that time, computations had been carried out using common fractions, although for geometric and astronomical computations, the sexagesimal system inherited from the Middle East was also used. It was through contacts with the Muslim culture that Europeans became familiar with the decimal place-value system, and such mathematicians as Gerbert of Aurillac encouraged the use of the new numbers in connection with the abacus. In the thirteenth century Leonardo of Pisa also helped to introduce this system of calculation into Europe, and in 1478 an arithmetic was published in Treviso, Italy, explaining the use of Hindu-Arabic numerals and containing computations in the form shown in Fig. 3. In the sixteenth century many scholars, including Robert Recorde (1510-1558) in Britain and Adam Ries (1492-1559) in Germany, advocated the use of the Hindu-Arabic system and established it as a universal standard.

The system was elegantly explained by the Flemish mathematician and engineer Simon Stevin (1548-1620) in his 1585 book *De Thiende (Decimals)*. Stevin took only a few pages to explain, in essentially modern terms, how to add, subtract, multiply, and divide decimal numbers. He then showed the application of this method of computing in finding land areas and the volumes of wine vats. He wrote concisely, as he said, "because here we are writing for teachers, not students." His notation appears slightly odd, however, since he put a circled 0 where we now have the decimal point, and thereafter indicated the rank of each digit by a similarly encircled number. For example, he would write 13.4832 as $13 \textcircled{4} 1 \textcircled{8} 2 \textcircled{3} 2 \textcircled{4}$. Here is his explanation of the problem of expressing $0.07 \div 0.00004$:

When the divisor is larger [has more digits] than the dividend, we adjoin to the dividend as many zeros as desired or necessary. For example, if $7 \textcircled{2}$ is to be divided by $4 \textcircled{5}$, I place some 0s next to the 7, namely 7000. This number is then divided as above, as follows:

$$\begin{array}{r} \cancel{3} \quad \cancel{2} \\ 7 \quad \cancel{0} \quad \cancel{0} \quad \cancel{0} \quad (1 \quad 7 \quad 5 \quad 0 \quad \textcircled{0} \\ \underline{4 \quad \cancel{4} \quad \cancel{4} \quad \cancel{4}} \end{array}$$

Hence the quotient is $1750 \textcircled{0}$. [Gericke and Vogel, 1965, p. 19]

Except for the location of the digits and the cross-out marks, this notation is essentially what is now done by school children in the United States. In other countries—Russia, for example—the divisor would be written just to the right of the dividend and the quotient just below the divisor.

Stevin also knew what to do if the division does not come out even. He pointed out that when $4 \textcircled{1}$ is divided by $3 \textcircled{2}$, the result is an infinite succession of 3s and that the exact answer will never be reached. He commented, "In such a case, one may go as far as the particular case requires and neglect the excess. It is certainly true that $13 \textcircled{0} 3 \textcircled{1} 3 \frac{1}{3} \textcircled{2}$, or $13 \textcircled{0} 3 \textcircled{1} 3 \textcircled{2} 3 \frac{1}{3} \textcircled{3}$, and so on are exactly equal to the required result, but our goal is to work only with whole numbers in this decimal computation, since we have in mind what occurs in human business, where [small parts of small measures] are ignored." Here we have a clear case in which the existence of infinite decimal expansions is admitted, without any hint of the possibility of irrational numbers. Stevin was an engineer, not a theoretical mathematician. His examples were confined to what is of practical value in business

and engineering, and he made no attempt to show how to calculate with an actually infinite decimal expansion.

Stevin did, however, suggest a reform in trigonometry that was ignored until the advent of hand-held calculators, remarking that, "if we can trust our experience (with all due respect to Antiquity and thinking in terms of general usefulness), it is clear that the series of divisions by 10, not by 60, is the most efficient, at least among those that are by nature possible." On those grounds, Stevin suggested that degrees be divided into decimal fractions rather than minutes and seconds. Modern hand-held calculators now display angles in exactly this way, despite the scornful remark of a twentieth-century mathematician that "it required four millennia to produce a system of angle measurement that is completely absurd."

8. The value of calculation

One cannot help noticing, alongside a few characteristics that are unique to a given culture, a large core of commonality in all this elementary mathematics. All of the treatises we have looked at pose problems of closely similar structure. This commonality is so great that any textbook of arithmetic published in the modern world up to very recent times is almost certain to repeat problems from the *Jiu Zhang Suanshu* or the Ahmose Papyrus or the *Brahmasphutasiddhanta* almost word for word. Thus, where Brahmagupta instructs the reader to "multiply the fruit and the requisition and divide by the argument in order to obtain the produce," Greenleaf (1876, p. 233) tells the reader to "find the required term by dividing the product of the second and third terms by the first." As far as clarity of exposition is concerned, one would have to give the edge to Brahmagupta. The number of ways of solving a mathematical problem is, after all, quite small; it is not surprising if two people in widely different circumstances come to the same conclusion.

Of course, when looking at the history of mathematics in the late nineteenth century, we tend to focus on the new research occurring at that time and overlook mere expositions of long-known mathematics such as one finds in the book of Greenleaf. But what Greenleaf was expounding was a set of mathematical skills that had been useful for many centuries. Like the *Jiu Zhang Suanshu* and the *Sun Zi Suan Jing*, his book contains discussions of the relations of various units of measure to one another and a large number of examples, both realistic and fanciful, showing how to carry out all the elementary operations. His job, in fact, was harder than that of the earlier authors, since he had to explain the relation between common fractions and decimal fractions, exchange rates for different kinds of currency, and many principles of commercial and inheritance law. If we now tend to regard such books as being of secondary importance in the history of mathematics, that is only because such a high superstructure has been built on that foundation. When we read the classics from Egypt, India, China, and Mesopotamia, on the other hand, we are looking at the frontier of knowledge in their time. It is a tribute to the authors of the treatises discussed in this chapter that they worked out and explained in clear terms a set of useful mathematical skills and bequeathed it to the world. For many centuries it could be said that the standard mathematical curriculum had a permanent value. Only in very recent years have the computational skills needed in commerce and law been superseded by the higher-level skills needed for deciding *when* and *what* to compute and how to interpret the results.

9. Mechanical methods of computation

Any study of the history of calculation must take account of the variety of computing hardware that people have invented and the software algorithms that are developed from time to time. In ancient China the software (decimal place-value system) was so good that the hardware (counting rods, counting boards, and abacus) worked with it very smoothly. The Greek and Roman system of writing numbers, however, was not a good representation of the decimal system, and the abacus was probably an essential tool of computation. When the graphical methods associated with Hindu-Arabic numerals were introduced into Europe, they were thought to be superior to the abacus.

9.1. Software: prosthaphæresis and logarithms. The graphic arithmetic that had vanquished the counting board a few centuries earlier still had certain laborious aspects connected with multiplication and division, which mathematicians kept trying to simplify. Consider, for example, the two three-digit numbers 476 and 835. To add these numbers we must perform three simple additions, plus two more that result from "carrying," a total of eight simple additions. In general, at most $3n - 1$ simple additions with $n - 1$ carryings will be required to add two n -digit numbers. Similarly, subtracting these numbers will require at most two borrowings, with consequent modification of the digits borrowed from, and three simple subtractions. For an n -digit number that is at most n simple subtractions and $n - 1$ borrowings.

On the other hand, to multiply two three-digit numbers will require nine simple multiplications followed by addition of the partial products, which will involve up to 10 more simple additions if carrying is involved. Thus we are looking at considerably more labor, with a number of additions and multiplications on the order of $2n^2$ if the two numbers each have n digits. Not only is a greater amount of time and effort needed, the procedure is obviously more error-prone. On the other hand, in a practical application in which we are multiplying, say, two seven-digit numbers (which would involve more than 100 simple multiplications and additions), we seldom need all 14 or 15 digits of the result. If we could improve the speed of the operation at the expense of some precision, the trade-off would be worthwhile.

Prosthaphæresis. The increased accuracy of astronomical instruments, among other applications, led to a need to multiply numbers having a large number of digits. As just pointed out, the amount of labor involved in multiplying two numbers increases as the product of the numbers of digits, while the labor of adding increases according to the number of digits in the smaller number. Thus, multiplying two 15-digit numbers requires over 200 one-digit multiplications, while adding the two numbers requires only 15 such operations (not including carrying). It was the large number of digits in the table entries that caused the problem in the first place, but the key to the solution turned out to be in the structural properties of sines and cosines. The process was called *prosthaphæresis*, from two Greek prefixes *pros-*, meaning *toward*, and *apo-*, meaning *from*, together with the root verb *haîrō*, meaning *I seize* or *I take*. Together these parts mean simply *addition and subtraction*.

There are hints of this process in several sixteenth-century works, but we shall quote just one example. In his *Trigonometria*, first published in Heidelberg in 1595, the theologian and mathematician Bartholomeus Pitiscus (1561-1613) posed the following problem: *to solve the proportion in which the first term is the radius, while the second and third terms are sines, avoiding multiplication and division.*

The problem here is to find the fourth proportional x , satisfying $r : a = b : x$, where r is the radius of the circle and a and b are two sines (half-chords) in the circle. We can see immediately that $x = ab/r$, but as Pitiscus says, the idea is to avoid the multiplication and division, since in the trigonometric tables of the time a and b might easily have seven or eight digits each.

The key to prosthaphæresis is the well-known formula

$$\sin \alpha \cos \beta = \frac{\sin(\alpha + \beta) + \sin(\alpha - \beta)}{2}.$$

This formula is applied as follows: If you have to multiply two large numbers, regard one of them as the sine of an angle, the other as the cosine of a second angle. (Since Pitiscus had only tables of sines, he had to use the complement of the angle having the second number as a sine.) Add the angles and take the sine of their sum to obtain the first term; then subtract the angles and take the sine of their difference to obtain a second term. Finally, divide the sum of the two terms by 2 to obtain the product. To take a very simple example, suppose that we wish to multiply 155 by 36. A table of trigonometric functions shows that $\sin 8^\circ 55' = 0.15500$ and $\cos 68^\circ 54' = 0.36000$. Hence

$$36 \cdot 155 = 10^5 \frac{\sin 77^\circ 49' + \sin(-59^\circ 59')}{2} = \frac{97748 - 86588}{2} = 5580.$$

In general, some significant figures will be lost in this kind of multiplication. For large numbers this procedure saves labor, since multiplying even two seven-digit numbers would tax the patience of most modern people. A further advantage is that prosthaphæresis is less error-prone than multiplication. Its advantages were known to the Danish astronomer Tycho Brahe (1546–1601), who used it in the astronomical computations connected with the extremely precise observations he made at his observatory during the latter part of the sixteenth century.

Logarithms. The problem of simplifying laborious multiplications, divisions, root extractions, and the like, was being attacked at the same time in another part of the world and from another point of view. The connection between geometric and arithmetic proportion had been noticed earlier by Chuquet, but the practical application of this fact had never been worked out. The Scottish laird John Napier, Baron of Murchiston (1550–1617), tried to clarify this connection and apply it. His work consisted of two parts, a theoretical part based on a continuous geometric model, and a computational part, involving a discrete (tabular) approximation of the continuous model. The computational part was published in 1614. However, Napier hesitated to publish his explanation of the theoretical foundation. Only in 1619, two years after his death, did his son publish the theoretical work under the title *Mirifici logarithmorum canonis descriptio* (*A Description of the Marvelous Rule of Logarithms*). The word *logarithm* means *ratio number*, and it was from the concept of ratios (geometric progressions) that Napier proceeded.

To explain his ideas Napier used the concept of moving points. He imagined one point P moving along a straight line from a point T toward a point S with decreasing velocity such that the ratio of the distances from the point P to S at two different times depends only on the difference in the times. (Actually, he called the line ending at S a sine and imagined it shrinking from its initial size TS , which he called the radius.) A second point is imagined as moving along a second line at a constant velocity equal to that with which the first point began. These two motions can be clarified by considering Fig. 4.

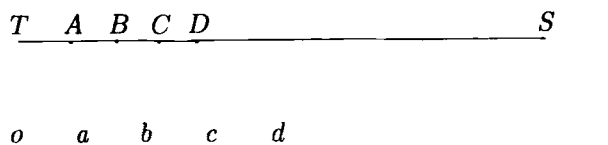


FIGURE 4. Geometric basis of logarithms.

The first point sets out from T at the same time and with the same speed with which the second point sets out from o . The first point, however slows down, while the second point continues to move at constant speed. The figure shows the locations reached at various times by the two points: When the first point is at A , the second is at a ; when the first point is at B , the second is at b ; and so on. The point moving with decreasing velocity requires a certain amount of time to move from T to A , the same amount of time to move from A to B , from B to C , and from C to D . Consequently, $TS : AS = AS : BS = BS : CS = CS : DS$.

The first point will never reach S , since it keeps slowing down, and its velocity at S would be zero. The second point will travel indefinitely far, given enough time. Because the points are in correspondence, the division relation that exists between two positions in the first case is mirrored by a subtractive relation in the corresponding positions in the second case. Thus, this diagram essentially changes division into subtraction and multiplication into addition. The top scale in Fig. 4 resembles a slide rule, and this resemblance is not accidental: a slide rule is merely an analog computer that incorporates a table of logarithms.

Napier's definition of the logarithm can be stated in the modern notation of functions by writing $\log(AS) = oa$, $\log(BS) = ob$, and so on; in other words, the logarithm increases as the "sine" decreases. These considerations contain the essential idea of logarithms. The quantity Napier defined is not the logarithm as we know it today. If points T , A , and P correspond to points o , a , and p , then

$$\overline{op} = \overline{oa} \log_k \left(\frac{\overline{PS}}{\overline{TS}} \right),$$

where $k = \overline{AS}/\overline{TS}$.

Arithmetical implementation of the geometric model. The geometric model just discussed is theoretically perfect, but of course one cannot put the points on a line into a table of numbers. It is necessary to construct the table from a finite set of points; and these points, when converted into numbers, must be rounded off. Napier was very careful to analyze the maximum errors that could arise in constructing such a table. In terms of Fig. 4, he showed that oa , which is the logarithm of AS , satisfies

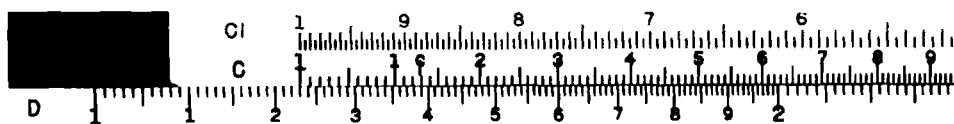
$$TA < oa < TA \left(1 + \frac{TA}{AS} \right).$$

(These inequalities are simple to prove, since the point describing oa has a velocity larger than the velocity of the point describing TA but less than TS/AS times the velocity of that point.) Thus, the tabular value for the logarithm of AS can be taken as the average of the two extremes, that is, $TA[1 + (TA/2AS)]$, and the relative error will be very small when TA is small.

Napier's death at the age of 67 prevented him from making some improvements in his system, which are sketched in an appendix to his treatise. These

improvements consist of scaling in such a way that the logarithm of 1 is 0 and the logarithm of 10 is 1, which is the basic idea of what we now call *common logarithms*. These further improvements to the theory of logarithms were made by Henry Briggs (1561–1630), who was in contact with Napier for the last two years of Napier's life and wrote a commentary on the appendix to Napier's treatise. As a consequence, logarithms to base 10 came to be known as *Briggsian logarithms*.

9.2. Hardware: slide rules and calculating machines. The fact that logarithms change multiplication into addition and that addition can be performed mechanically by sliding one ruler along another led to the development of rulers with the numbers arranged in proportion to their logarithms (slide rules). When one such scale is slid along a second, the numbers pair up in proportion to the distance slid, so that if 1 is opposite 5, then 3 will be opposite 15. Multiplication and division are then just as easy to do as addition and subtraction would be. The process is the same for both multiplication and division, as it was in the Egyptian graphical system, which was also based on proportion. Napier himself designed a system of rods for this purpose. A variant of this linear system was a system of sliding circles. Such a circular slide rule was described in a pamphlet entitled *Grammelogia* written in 1630 by Richard Delamain (1600–1644), a mathematics teacher living in London. Delamain urged the use of this device on the grounds that it made it easy to compute compound interest. Two years later the English clergyman William Oughtred (1574–1660) produced a similar description of a more complex device. Oughtred's *circles of proportion*, as he called them, gave sines and tangents of angles in various ranges on eight different circles. Because of their portability, slide rules remained the calculating machine of choice for engineers for 350 years, and improvements were still being made in them in the 1950s. Different types of slide rule even came to have different degrees of prestige, according to the number of different scales incorporated into them.



Portions of the C, D, and CI scales of a slide rule. Adjacent numbers on the C and D scales are in proportion, so that $1 : 1.23 :: 1.3 : 1.599 :: 1.9 : 2.337$. Thus, the position shown here illustrates the multiplication $1.23 \cdot 1.3 = 1.599$, the division $1.722 \div 1.4 = 1.23$, and many other computations. Some visual error is inevitable. The CI (inverted) scale gives the reciprocals of the numbers on the C scale, so that division can be performed as multiplication, only using the CI scale instead of the C scale. Decimal points have to be provided by the user.

Slide rule calculations are floating-point numbers with limited accuracy and necessary round-off error. When computing with integers, we often need an exact answer. To achieve that result, adding machines and other digital devices have been developed over the centuries. An early design for such a device with a series of interlocking wheels can be found in the notebooks of Leonardo da Vinci (1452–1519). Similar machines were designed by Blaise Pascal (1623–1662) and Gottfried Wilhelm Leibniz (1646–1716). Pascal's machine was a simple adding machine that

depended on turning a crank a certain number of times in order to find a sum. Leibniz used a variant of this machine with a removable set of wheels that would multiply, provided that the user kept count of the number of times the crank was turned.

Machines designed to calculate for a specific purpose continued to be built for centuries, but all were doomed to be replaced by the general-purpose information processor that has spread like wildfire around the world in the past two decades. The first prefiguration of such a machine was Charles Babbage's difference engine, designed in the 1830s but built only partially many decades later. Although only one such machine seems to have been built, and only part of Babbage's more ambitious analytical engine was constructed, the idea of a general-purpose computer that could accept instructions and modify its operation in accordance with them was a brilliant innovation. Unfortunately, the full implementation of this idea could not be carried out by mechanical devices with moving parts. It needed the reliability of electronics, first thermionic valves (vacuum tubes) and then transistors, to produce the marvels of technology that we all use nowadays. That technology was developed in Britain and the United States, greatly stimulated by the needs of code breaking during World War II.

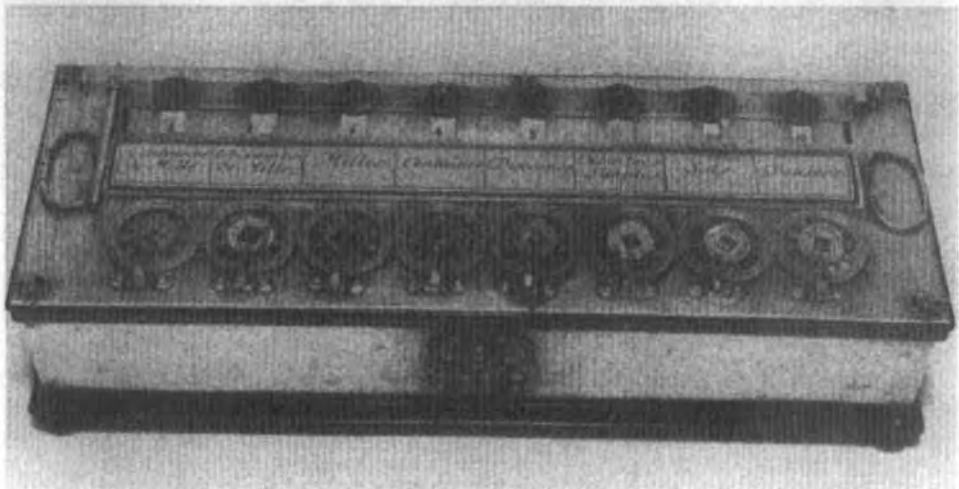
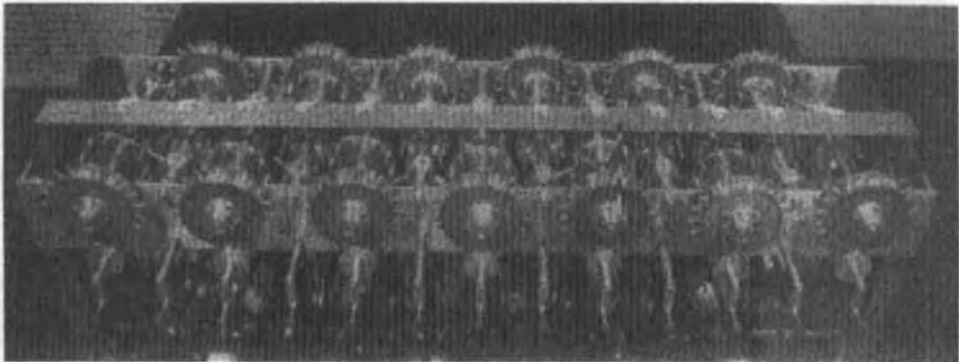
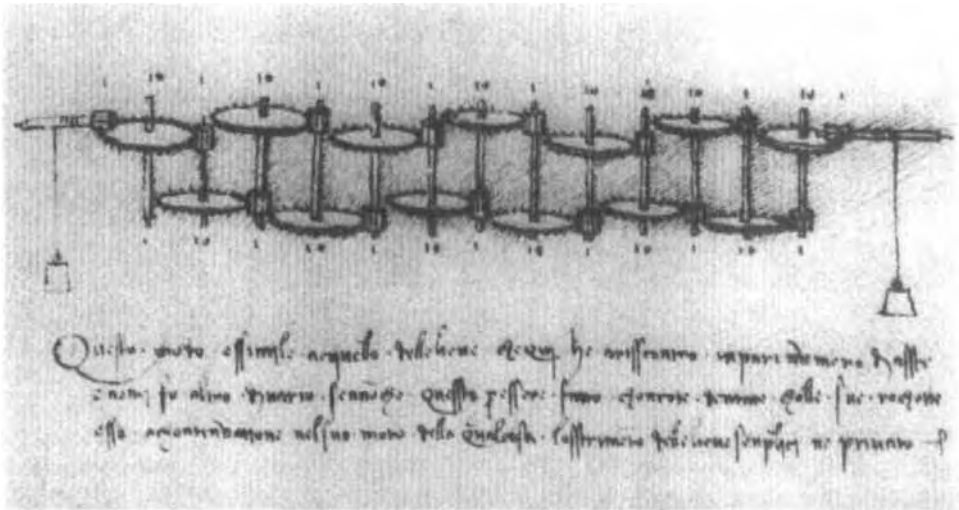
Meanwhile, fixed-purpose machines that could perform only a limited number of set arithmetic operations, continued to be built and improved upon. In the late days of World War II the "latest thing" in automated calculation was a machine with many fragile moving parts.

Hand-held calculators, at first very limited, began to supplant the slide rule during the 1970s. They were easier to use, infinitely faster, and much more accurate than slide rules, and they soon became much cheaper. The only disadvantage they had in comparison with slide rules was in durability.⁶ A popular American textbook of college algebra published in 1980 weighed the advantages and disadvantages of slide rules, calculators, and tables of logarithms, and summed up for the jury in favor of using tables of logarithms, which had the cheapness of slide rules but more accuracy. Even that recently it was an extremely refined and expensive hand calculator that had more than a few dozen memory cells. It was not possible to foresee the explosion of computing power that was to result from the development of methods of producing huge quantities of memory on tiny chips at extremely low prices.

For several decades following World War II there were two types of calculating devices: Slide rules and cheap adding machines served the individual; more expensive calculators and the early gigantic computers such as ENIAC (Electronic Numerical Integrator And Computer) were used by large corporations. The overwhelming penetration of modern computers into nearly every human activity, especially their use for word processing and graphics, is due to the vision of people such as Charles Babbage, who realized that they must be able to use Boolean logic in addition to their calculating capacity.

Two mathematicians figure prominently in the development of this vision of the computer. One was Alan Turing (1912-1954), a British mathematician whose 1937 paper "On computable numbers" contained the idea of a universal computer

⁶ One might think that there would be no further market for slide rules. To the contrary, an entire website is devoted to buying and selling them. The webmaster points out to the site visitor that the computer on which the purchase is being made will be in a landfill 50 years hence, whereas the slide rule will be only well broken in.



Top: Leonardo da Vinci's design for an adding machine. Center: model of the machine. Bottom: Pascal's adding machine. Courtesy of IBM Corporate Images.



The future of calculation, as depicted in the April 17, 1944 issue of *Newsweek*. The potential of mechanical computing devices is limited by wear and tear on its moving parts. A slide rule suffers very little from such wear, but the more sophisticated adding machines of Pascal had to be designed with counterweights to delay the inevitable lopsidedness that results from wear. Courtesy of the Bodine Electric Company, Chicago.

now known as a *Turing machine*. The crucial element was programmability: Turing envisioned a tape filled with information that could be fed into the machine. As the machine read the information, it would modify its own internal state and then move ahead or back to an adjacent instruction. In this way, Turing went a step

further than Babbage had envisioned. Babbage's assistant Augusta Ada Lovelace (1815–1852) had written that Babbage's analytical engine could do only what it was told to do, but Turing believed that the difference between human intelligence and a computer was not so stark as that. He considered it possible that what appeared as human creativity might be the result of some information delivered at an earlier time, and that computers might mimic this apparent creativity.

The other mathematician was John von Neumann (1903–1957), who was at Princeton in 1936 when Turing came there as a graduate student.⁷ Von Neumann became involved in the development of the computer while working as a consultant at the Aberdeen Proving Grounds in Maryland. There, in August 1944, he met Herman H. Goldstine (1913–2004), who told him about ENIAC. From our perspective, two generations on, ENIAC looks like the brontosaurus of computing technology. Here is a description of it from a website devoted to its history:⁸

When it was finished, the ENIAC filled an entire room, weighed thirty tons, and consumed two hundred kilowatts of power. It generated so much heat that it had to be placed in one of the few rooms at the University [of Pennsylvania] with a forced air cooling system. Vacuum tubes, over 19,000 of them, were the principal elements in the computer's circuitry. It also had fifteen hundred relays and hundreds of thousands of resistors, capacitors, and inductors. All of this electronics was held in forty-two panels nine feet tall, two feet wide, and one foot thick.

Despite its size, ENIAC had very little memory—only 1000 bits of RAM! More-over the use of vacuum tubes (thermionic valves) meant frequent breakdowns—on every 8 minutes on the average, until the operators reduced the voltage and current to the minimum; after that, breakdowns occurred only once every two days on the average. Most seriously, it could not be programmed in the present-day sense of the word. It simply had to be set up for each particular computation. Von Neumann and the builders of ENIAC collaborated on the construction of EDVAC (Electronic Discrete Variable Automatic Computer), which had the ability to read stored programs. Of course, those programs had to be written in machine language, a serious drawback for human interaction, but von Neumann's basic idea was sound. We have engineers to thank that the vacuum tube has been replaced by the transistor, that transistors can now be etched onto tiny computer chips, and that production methods have made it possible to produce for a very modest price the machines that everyone now uses with such ease.

9.3. The effects of computing power. A crystal ball can be very cloudy, even in relation to the eternal truths of mathematics. A book of mathematical tables and formulas (Burlington, 1958) purchased by the author nearly half a century ago confidently assured its readers in a note from the publishers that

the subject matter [of this book] is not ephemeral but everlasting—
as true in the future as it has been in the past. By all means, retain

⁷ According to Heppenheimer (1990), von Neumann offered Turing a position as his assistant, but Turing preferred to return to Cambridge.

⁸ <http://ei.cs.vt.edu/~history/ENIAC.Richey.HTML>.

this book for your own reference library. You will need it many times in years to come.

That book remains on the author's shelf, unopened since about 1985. The publishers in their confidence had overlooked the fact that the eternal truths of mathematics need not be reconstructed every time they are needed. Machines can store them and do the unimaginative computational work more efficiently and accurately than people can.

One result of all this magnificent computer engineering is that mathematics education faces a dilemma. On the one hand, the skills involved in doing elementary arithmetic, algebra, and calculus are now as obsolete as the skill of writing a letter in longhand. What is the point of teaching students how to solve quadratic equations, factor polynomials, carry out integration by parts, and solve differential equations when readily available programs such as *Mathematica*, *Maple*, *Matlab*, and others can produce the result in a split second with guaranteed accuracy? On the other hand, solving mathematical problems requires quantitative reasoning, and no one has yet found any way to teach quantitative reasoning without assuming a familiarity with these basic skills. How can you teach what multiplication is without making students learn the multiplication table? How can you explain the theory of equations without making students solve a few equations? If mathematics education is to be in any way relevant to the lives of the students who are its clients, it must be able to explain in cogent terms the reason for the skills it asks them to undergo so much boredom to learn, or else find other skills to teach them.

Questions and problems

6.1. Double the hieroglyphic number $\begin{array}{c} ||| \\ ||| \end{array} \begin{array}{c} \cap \\ \cap \cap \end{array}$.

6.2. Multiply 27 times 42 the Egyptian way.

6.3. (Stated in the Egyptian style.) Calculate with 13 so as to obtain 364.

6.4. Problem 23 of the Ahmose Papyrus asks what parts must be added to the sum of $\bar{4}$, $\bar{8}$, $\bar{10}$, $\bar{30}$, and $\bar{45}$ to obtain $\bar{3}$. See if you can obtain the author's answer of $\bar{9} \bar{40}$, starting with his technique of magnifying the first row by a factor of 45. Remember that $\frac{5}{8}$ must be expressed as $\bar{2} \bar{8}$.

6.5. Problem 24 of the Ahmose Papyrus asks for a number that yields 19 when its seventh part is added to it, and concludes that one must perform on 7 the same operations that yield 19 when performed on 8. Now in Egyptian terms, 8 must be multiplied by $2 \bar{4} \bar{8}$ in order to obtain 19. Multiply this number by 7 to obtain the scribe's answer, $16 \bar{2} \bar{8}$. Then multiply that result by $\bar{7}$, add the product to the result itself, and verify that you do obtain 19, as required.

6.6. Problem 33 of the Ahmose Papyrus asks for a quantity that yields 37 when increased by its two parts (two-thirds), its half, and its seventh part. Try to get the author's answer: The quantity is $16 \bar{56} \bar{679} \bar{776}$. [Hint: Look in the table of doubles of parts for the double of $\bar{97}$. The scribe first tried the number 16 and found that the result of these operations applied to 16 fell short of 37 by the double of $\bar{42}$, which, as it happens, is exactly $1 \bar{3} \bar{2} \bar{7}$ times the double of $\bar{97}$.]

6.7. Verify that the solution to Problem 71 ($2\overline{3}$) is the correct *pesu* of the diluted beer discussed in the problem.

6.8. Compare the *pesu* problems in the Ahmose Papyrus with the following problem, which might have been taken from almost any algebra book written in the past century: *A radiator is filled with 16 quarts of a 10% alcohol solution. If it requires a 30% alcohol solution to protect the radiator from freezing when it is turned off, how much 95% solution must be added (after an equal amount of the 10% solution is drained off) to provide this protection?* Think of the alcohol as the grain in beer and the liquid in the radiator as the beer. The liquid has a *pesu* of 10. What is the *pesu* that it needs to have, and what is the *pesu* of the liquid that is to be used to achieve this result?

6.9. Verify that the solution $5\overline{10}$ given above for Problem 35 is correct, that is, multiply this number by 3 and by $\overline{3}$ and verify that the sum of the two results is 1.

6.10. Why do you suppose that the author of the Ahmose Papyrus did not choose to say that the double of the thirteenth part is the seventh part plus the ninety-first part, that is,

$$\frac{2}{13} = \frac{1}{7} + \frac{1}{91}?$$

Why is the relation

$$\frac{2}{13} = \frac{1}{8} + \frac{1}{52} + \frac{1}{104}$$

made the basis for the tabular entry instead?

6.11. Generalizing Question 6.10, investigate the possibility of using the identity

$$\frac{2}{p} = \frac{1}{\left(\frac{p+1}{2}\right)} + \frac{1}{p\left(\frac{p+1}{2}\right)}$$

to express the double of the reciprocal of an odd number p as a sum of two reciprocals. Which of the entries in the table of Fig. 1 can be obtained from this pattern? Why was it not used to express $\frac{2}{15}$?

6.12. Why not simply write $\overline{13}\overline{13}$ to stand for what we call $\frac{2}{13}$? What is the reason for using two or three other “parts” instead of these two obvious parts?

6.13. Could the ability to solve a problem such as Problem 35, discussed in Subsection 1.2 of this chapter, have been of any practical use? Try to think of a situation in which such a problem might arise.

6.14. We would naturally solve many of the problems in the Ahmose Papyrus using an equation. Would it be appropriate to say that the Egyptians solved equations, or that they did algebra? What does the word *algebra* mean to you? How can you decide whether you are performing algebra or arithmetic?

6.15. Why did the Egyptians usually begin the process of division by multiplying by $\overline{3}$ instead of the seemingly simpler $\overline{2}$?

6.16. Early mathematicians must have been adept at thinking in terms of expressions. But considering the solutions to the riders-and-carts problem and the colorful language of Brahmagupta in relation to the Rule of Three, one might look at the situation from a different point of view. Perhaps these early mathematicians were good “dramatists.” In any algorithm the objects we now call variables amount to special “roles” played, with different numbers being assigned to “act” in those roles;

an algorithm amounts to the drama that results when these roles are acted. That is why it is so important that each part of the algorithm have its own name. The letters that we use for variables amount to names assigned to roles in the drama. A declaration of variables at the beginning of a program is analogous to the section that used to be titled “Dramatis Personæ” at the beginning of a play.

Explain long division from this point of view, using the roles of dividend, divisor, quotient, and remainder.

6.17. Imitate the reasoning used in solving the problem of riders and carts above to solve Problem 17 of the *Sun Zi Suan Jing*. The problem asks how many guests were at a banquet if every two persons shared a bowl of rice, every three persons a bowl of soup, and every four persons a bowl of meat, leading to a total of 65 bowls. Don’t use algebra, but try to explain the rather cryptic solution given by Sun Zi: Put down 65 bowls, multiply by 12 to obtain 780, and divide by 13 to get the answer.

6.18. Compare the following loosely interpreted problems from the *Jiu Zhang Suanshu* and the Ahmose Papyrus. First, from the *Jiu Zhang Suanshu*: Five officials went hunting and killed five deer. Their ranks entitle them to shares in the proportion 1 : 2 : 3 : 4 : 5. What part of a deer does each receive?

Second, from the Ahmose Papyrus (Problem 40): 100 loaves of bread are to be divided among five people (in arithmetic progression), in such a way that the amount received by the last two (together) is one-seventh of the amount received by the first three (together). How much bread does each person receive?

6.19. Compare the interest problem (Problem 20 of Chapter 3) from the *Jiu Zhang Suanshu* discussed above, with the following problem, taken from the American textbook *New Practical Arithmetic* by Benjamin Greenleaf (1876):

The interest on \$200 for 4 months being \$4, what will be the interest on \$590 for 1 year and 3 months?

Are there any significant differences at all in the nature of the two problems, written nearly 2000 years apart?

6.20. Problem 4 in Chapter 6 of the *Jiu Zhang Suanshu* involves what is called *double false position*. The problem reads as follows: A number of families contribute equal amounts to purchase a herd of cattle. If the contribution (the same for each family) were such that seven families contribute a total of 190 [units of money], there would be a deficit of 330 [units of money]; but if the contribution were such that nine families contribute 270 [units of money], there would be a surplus of 30 [units of money]. Assuming that the families each contribute the correct amount, how much does the herd cost, and how many families are involved in the purchase? Explain the solution given by the author of the *Jiu Zhang Suanshu*, which goes as follows. Put down the proposed values (assessment to each family, that is, $\frac{190}{7}$ and $\frac{270}{9} = 30$), and below each put down the corresponding surplus or deficit (a positive number in each case). Cross-multiply and add the products to form the *shi* ($30 \cdot \frac{190}{7} + 330 \cdot 2709 = \frac{75000}{7}$). Add the surplus and deficit to form the *fa* ($330 + 30 = 360$). Subtract the smaller of the proposed values from the larger, to get the difference ($\frac{270}{9} - \frac{190}{7} = \frac{20}{7}$). Divide the *shi* by the difference to get the cost of the goods ($\frac{75000}{20} = 3750$); divide the *fa* by the difference to get the number of families ($\frac{360}{20/7} = 126$).

6.21. Compare the pond-filling problem (Problem 26 of Chapter 6) of the *Jiu Zhang Suanshu* (discussed above) with the following problem from Greenleaf (1876, p. 125): *A cistern has three pipes; the first will fill it in 10 hours, the second in 15 hours, and the third in 16 hours. What time will it take them all to fill it?* Is there any real difference between the two problems?

6.22. The fair taxation problem from the *Jiu Zhang Suanshu* considered above treats distances and population with equal weight. That is, if the population of one county is double that of another, but that county is twice as far from the collection center, the two counties will have exactly the same tax assessment in grain and carts. Will this impose an equal burden on the taxpayers of the two counties? Is there a direct proportionality between distance and population that makes them interchangeable from the point of view of the taxpayers involved? Is the growing of extra grain to pay the tax fairly compensated by a shorter journey?

6.23. Perform the division $\frac{980}{45}$ following the method used by Brahmagupta.

6.24. Convert the sexagesimal number 5; 35, 10 to decimal form and the number 314.7 to sexagesimal form.

6.25. As mentioned in connection with the lunisolar calendar, 19 solar years equal almost exactly 235 lunar months. (The difference is only about two hours.) In the Julian calendar, which has a leap year every fourth year, there is a natural 28-year cycle of calendars. The 28 years contain exactly seven leap-year days, giving a total of exactly 1461 weeks. These facts conjoin to provide a natural 532-year cycle ($532 = 28 \cdot 19$) of calendars incorporating the phases of the Moon. In particular, Easter, which is celebrated on the Sunday after the first full Moon of spring, has a 532-year cycle (spoiled only by the two-hour discrepancy between 19 years and 235 months). According to Simonov (1999), this 532-year cycle was known to Cyrus (Kirik) of Novgorod when he wrote his “Method by which one may determine the dates of all years” in the year 6644 from the creation of the world (1136 CE). Describe how you would create a table of dates of Easter that could, in principle, be used for all time, so that a user knowing the number of the current year could look in the table and determine the date of Easter for that year. How many rows and how many columns should such a table have, and how would it be used?

6.26. From 1901 through 2099 the Gregorian calendar behaves like the Julian calendar, with a leap year every four years. Hence the 19-year lunar cycle and 28-year cycle of days interact in the same way during these two centuries. As an example, we calculate the date of Easter in the year 2039. The procedure is first to compute the remainder when 2039 is divided by 19. The result is 6 ($2039 = 19 \times 107 + 6$). This number tells us where the year 2039 occurs in the 19-year lunar cycle. In particular, by consulting the table below for year 6, we find that the first full Moon of spring in 2039 will occur on April 8. (Before people became familiar with the use of the number 0, it was customary to add 1 to this remainder, getting what is still known in prayer books as the *golden number*. Thus the golden number for the year 2039 is 7.)

We next determine by consulting the appropriate calendar in the 28-year cycle which day of the week April 8 will be. In fact, it will be a Friday in 2039, so that Easter will fall on April 10 in that year. The dates of the first full Moon in spring for the years of the lunar cycle are as follows. The year numbers are computed as above, by taking the remainder when the Gregorian year number is divided by 19.

Year	0	1	2	3	4	5	6
Full Moon	Apr. 14	Apr. 3	Mar. 23	Apr. 11	Mar. 31	Apr. 18	Apr. 8
Year	7	8	9	10	11	12	13
Full Moon	Mar. 28	Apr. 16	Apr. 5	Mar. 25	Apr. 13	Apr. 2	Mar. 22
Year	14	15	16	17	18		
Full Moon	Apr. 10	Mar. 30	Apr. 17	Apr. 7	Mar. 27		

Using this table, calculate the date of Easter for the years from 2040 through 2045. You can easily compute the day of the week for each of these dates in a given year, starting from the fact that March 21 in the year 2000 was a Tuesday. [Note: If the first full Moon of spring falls on a Sunday, Easter is the following Sunday.]

6.27. Prosthaphæresis can be carried out using only a table of cosines by making use of the formula

$$\cos \alpha \cos \beta = \frac{\cos(\alpha + \beta) + \cos(\alpha - \beta)}{2}.$$

Multiply 3562 by 4713 using this formula and a table of cosines. (It is fair to use your calculator as a table of cosines; just don't use its arithmetical capabilities.)

6.28. Do the multiplication 742518 · 635942 with pencil and paper without using a hand calculator, and time yourself. Also count the number of simple multiplications you do. Then get a calculator that will display 12 digits and do the same problem on it to see what errors you made, if any. (The author carried out the 36 multiplications and 63 additions in just under 5 minutes, but had two digits wrong in the answer as a result of incorrect carrying.)

Next, do the same problem using prosthaphæresis. (Again, you may use your hand calculator as a trigonometric table.) How much accuracy can you obtain this way? With a five-place table of cosines, using interpolation, the author found the two angles to be 50.52° and 42.05°. The initial digits of the answer would thus be those of (cos(8.47°) + cos(92.57°))/2, yielding 47213 as the initial digits of the 12-digit number. On the other hand, using a calculator that displays 14 digits, one finds the angles to be 50.510114088363° and 42.053645425939°. That same calculator then returns all 12 digits of the correct answer as the numerical value of (cos(8.45646866242°) + cos(92.563759514302°))/2. Compared with the time to do the problem in full the time saved was not significant.

Finally, do the problem using logarithms. Again, you may use your calculator to look up the logarithms, since a table is probably not readily available.

CHAPTER 7

Ancient Number Theory

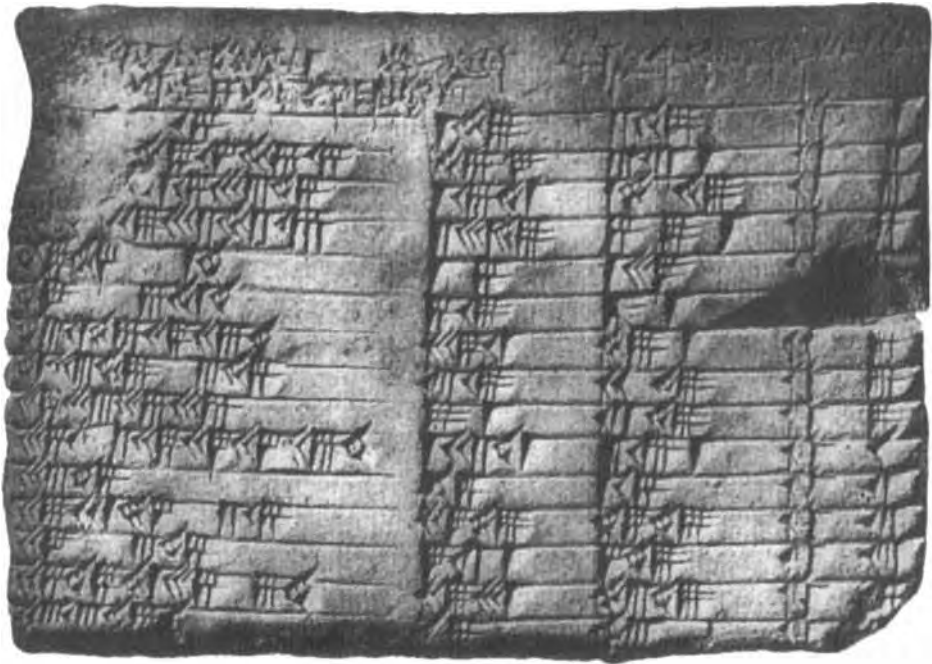
The impossibility of getting square roots to come out even, in connection with applications of the Pythagorean theorem, may have caused mathematicians to speculate on the difference between numbers that have (rational) square roots and those that do not. We shall take this problem as the starting point for our discussion of number theory, and we shall see two responses to this problem: first, in the present chapter, to find out when indeterminate quadratic equations have rational solutions; second, in Chapter 8, to create new numbers to play the role of square roots when no rational square root exists.

1. Plimpton 322

Rational numbers satisfying a quadratic equation are at the heart of a cuneiform tablet from the period 1900–1600 BCE, number 322 of the Plimpton collection at Columbia University. The numbers on this tablet have intrigued many mathematically oriented people, leading to a wide variety of speculation as to the original purpose of the tablet. We are not offering any new conjectures as to that purpose here, only a discussion of some earlier ones.

As you can see from the photograph on p. 160, there are a few chips missing, so that some of the cuneiform numbers in the tablet will need to be restored by plausible conjecture. Notice also that the column at the right-hand edge contains the cuneiform numbers in the sequence 1, 2, 3, 4, ..., ..., 7, 8, 9, 10, 11, 12, 13, ..., Obviously, this column merely numbers the rows. The column second from the right consists of identical symbols that we shall ignore entirely. Pretending that this column is not present, if we transcribe only what we can see into our version of sexagesimal notation, denoting the chipped-off places with brackets ($\{\dots\}$), we get the four-column table shown below.

Before analyzing the mathematics of this table, we make one preliminary observation: Row 13 is anomalous, in that the third entry is smaller than the second entry. For the time being we shall ignore this row and see if we can figure out how to correct it. Since the long numbers in the first column must be the result of computation—it is unlikely that measurements could be carried out with such precision—we make the reasonable conjecture that the shorter numbers in the second and third columns are data. As mentioned in Chapter 6, the Mesopotamian mathematicians routinely associated with any pair of numbers (a, b) two other numbers: their average $(a + b)/2$ and their semidifference $(b - a)/2$. Let us compute these numbers for all the rows except rows 13 and 15, to see how they would have appeared to a mathematician of the time. We get the following 13 pairs of numbers, which we write in decimal notation: (144, 25), (7444, 4077), (5625, 1024), (15625, 2916), (81, 16), (400, 81), (2916, 625), (1024, 225), (655, 114), (6561, 1600), (60, 15), (2304, 625), (2500, 729).



Plimpton 322. © Rare Book and Manuscript Library, Columbia University.

	Width	Diagonal	
[...] 15	1,59	2,49	1
[...] 58,14,50,6,15	56,7	3,12,1	2
[...] 41,15,33,45	1,16,41	1,50,49	3
[...] 29,32,52,16	3,31,49	5,9,1	4
48,54,1,40	1,5	1,37	5
47,6,41,40	5,19	8,1	6
43,11,56,28,26,40	38,11	59,1	7
41,33,59,3,45	13,19	20,49	8
38,33,36,36	9,1	12,49	9
35,10,2,28,27,24,26	1,22,41	2,16,1	10
33,45	45	1,15	11
29,21,54,2,15	27,59	48,49	12
27,[...],3,45	7,12,1	4,49	13
25,48,51,35,6,40	29,31	53,49	14
23,13,46,40	[...]	[...]	[...]

You will probably recognize a large number of perfect squares here. Indeed, *all* of these numbers, except for those corresponding to rows 2, 9, and 11 are perfect squares: 10 pairs of perfect squares out of thirteen! That is too unusual to be a mere coincidence. A closer examination reveals that they are squares of numbers whose only prime factors are 2, 3, and 5. Now these are precisely the prime factors of the number 60, which the Mesopotamian mathematicians used as a base. That means that the reciprocals of these numbers will have terminating sexagesimal expansions.

We should therefore keep in mind that the reciprocals of these numbers may play a role in the construction of the table.

Notice also that these ten pairs are all *relatively prime* pairs. Let us now denote the square root of the average by p and the square root of the semidifference by q . Column 2 will then be $p^2 - q^2$, and column 3 will be $p^2 + q^2$. Having identified the pairs (p, q) as important clues, we now ask *which* pairs of integers occur here and how they are arranged. The values of q , being smaller, are easily handled. The smallest q that occurs is 5 and the largest is 54, which also is the largest number less than 60 whose only prime factors are 2, 3, and 5. Thus, we could try constructing such a table for all values of q less than 60 having only those prime factors. But what about the values of p ? Again, ignoring the rows for which we do not have a pair (p, q) , we observe that the rows occur in decreasing order of p/q , starting from $12/5 = 2.4$ and decreasing to $50/27 = 1.85185185\dots$. Let us then impose the following conditions on the numbers p and q :

1. The integers p and q are relatively prime.
2. The only prime factors of p and q are 2, 3, and 5.
3. $q < 60$.
4. $1.8 \leq p/q \leq 2.4$

Now, following an idea of Price (1964), we ask which possible (p, q) satisfy these four conditions. We find that every possible pair occurs with only five exceptions: (2, 1), (9, 5), (15, 8), (25, 12), and (64, 27). There are precisely five rows in the table—rows 2, 9, 11, 13, and 15—for which we did not find a pair of perfect squares. Convincing proof that we are on the right track appears when we arrange these pairs in decreasing order of the ratio p/q . We find that (2, 1) belongs in row 11, (9, 5) in row 15, (15, 8) in row 13, (25, 12) in row 9, and (64, 27) in row 2, precisely the rows for which we did not previously have a pair p, q . The evidence is overwhelming that these rows were intended to be constructed using these pairs (p, q) . When we replace the entries that we can read by the corresponding numbers $p^2 - q^2$ in column 2 and $p^2 + q^2$ in column 3, we find the following:

In row 2, the entry 3,12,1 has to be replaced by 1,20,25, that is, 11521 becomes 4825. The other entry in this row, 56,7, is correct.

In row 9, the entry 9,1 needs to be replaced by 8,1, so here the writer simply inserted an extra unit character.

In row 11, the entries 45 and 75 must be replaced by 3 and 5; that is, both are divided by 15. It has been remarked that if these numbers were interpreted as $45 \cdot 60$ and $75 \cdot 60$, then in fact, one would get $p = 60$, $q = 30$, so that this row was not actually “out of step” with the others. But of course when that interpretation is made, p and q are no longer relatively prime, in contrast to all the other rows.

In row 13 the entry 7,12,1 must be replaced by 2,41; that is, 25921 becomes 161. In other words, the table entry is the square of what it should be.

The illegible entries in row 15 now become 56 and 106. The first of these is consistent with what can be read on the tablet. The second appears to be 53, half of what it should be.

The final task in determining the mathematical meaning of the tablet is to explain the numbers in the first column and interpolate the missing pieces of that column. Notice that the second and third columns in the table are labeled “width”

and “diagonal.” Those labels tell us that we are dealing with dimensions of a rectangle here, and that we should be looking for its length. By the Pythagorean theorem, that length is $\sqrt{(p^2 + q^2)^2 - (p^2 - q^2)^2} = \sqrt{4p^2q^2} = 2pq$. Even with this auxiliary number, however, it requires some ingenuity to find a formula involving p and q that fits the entries in the first column that can be read. If the numbers in the first column are interpreted as the sexagesimal representations of numbers between 0 and 1, those in rows 5 through 14—the rows that can be read—all fit the formula¹

$$\left(\frac{p/q - q/p}{2}\right)^2.$$

Assuming this interpretation, since it works for the 10 entries we can read, we can fill in the missing digits in the first four and last rows. This involves adding two digits to the beginning of the first four rows, and it appears that there is just the right amount of room in the chipped-off place to allow this to happen. The digits that occur in the bottom row are 23,13,46,40, and they are consistent with the parts that can be read from the tablet itself.

The purpose of Plimpton 322: some conjectures. The *structure* of the tablet is no longer a mystery, unless one counts the tiny mystery of explaining the misprint in row 2, column 3. Its *purpose*, however, is not clear. What information was the table intended to convey? Was it intended to be used as people once used tables of products, square roots, and logarithms, that is, to look up a number or pair of numbers? If so, which columns contained the input and which the output? One geometric problem that can be solved by use of this tablet is that of multiplying a square by a given number, that is, given a square of side a , it is possible to find the side b of a square whose ratio to the first square is given in the first column. To do so, take a rope whose length equals the side a and divide it into the number of equal parts given in the second column, then take a second rope with the same unit of length and total length equal to the number of units in the third column and use these two lengths to form a leg and the hypotenuse of a right triangle. The other leg will then be the side of a square having the given ratio to the given square. The problem of shrinking or enlarging squares was considered in other cultures, but such an interpretation of Plimpton 322 has only the merit that there is no way of proving the tablet *wasn't* used in this way. There is no proof that the tablet was ever put to this use.

Friberg (1981) suggested that the purpose of the tablet was trigonometrical, that it was a table of squares of tangents. Columns 2 and 3 give one leg and the hypotenuse of 15 triangles with angles intermediate between those of the standard 45-45-90 and 30-60-90 triangles. What is very intriguing is that the table contains all possible triangles whose shapes are between these two and whose legs have lengths that are multiples of a standard unit by numbers having only 2, 3, and 5

¹ In some discussions of Plimpton 322 the claim is made that a sexagesimal 1 should be placed before each of the numbers in the first column. Although the tablet is clearly broken off on the left, it does not appear from pictures of the tablet—the author has never seen it “live”—that there were any such digits there before. Neugebauer (1952, p. 37) claims that parts of the initial 1 remain from line 4 on “as is clearly seen from the photograph” and that the initial 1 in line 14 is completely preserved. When that assumption is made, however, the only change in the interpretation is a trivial one: The negative sign in the formula must be changed to a positive sign, and what we are interpreting as a column of squares of tangents becomes a column of squares of secants, since $\tan^2 \theta + 1 = \sec^2 \theta$.

as factors. Of all right triangles, the 45-45-90 and the 30-60-90 are the two that play the most important role in all kinds of geometric applications; plastic models of them were once used as templates in mechanical drawing, and such models are still sold. It is easy to imagine that a larger selection of triangle shapes might have been useful in the past, before modern drafting instruments and computer-aided design. Using this table, one could build 15 model triangles with angles varying in increments of approximately 1° . One can imagine such models being built and the engineer of 4000 years ago reaching for a "number 7 triangle" when a slope of $574/675 = .8504$ was needed. However, this scenario still lacks plausibility. Even if we assume that the engineer kept the tablet around as a reference when it was necessary to know the slope, the tablet stores the *square* of the slope in column 1. It is difficult to imagine any engineering application for that number.

Having failed to find a geometric explanation of the tablet, we now explore possible associations of the tablet with Diophantine equations, that is, equations whose solutions are to be rational numbers, in this case numbers whose numerators and denominators are products of only the first three prime numbers. The left-hand column contains numbers that are perfect squares and remain perfect squares when 1 is added to them. In other words, it gives u^2 for solutions to the Diophantine equation $u^2 + 1 = v^2$. This equation was much studied in other cultures, as we shall see below. If the purpose of the table were to generate solutions of this equation, there would of course be no reason to give v^2 , since it could be obtained by placing a 1 before the entry in the first column. The use of the table would then be as follows: Square the entry in column 3, square the entry in column 2, then divide each by the difference of these squares. The results of these two divisions are v^2 and u^2 respectively. In particular, u^2 is in column 1. The numbers p and q that generate the two columns can be arbitrary, but in order to get a sexagesimally terminating entry in the first column, the difference $(p^2 + q^2)^2 - (p^2 - q^2)^2 = 4p^2q^2$ should have only 2, 3, and 5 as prime factors, and hence p and q also should have only these factors. Against this interpretation there lies the objection that p and q are concealed from the casual reader of the tablet. If the purpose of the tablet was to show how to generate u and v or u^2 and $v^2 = u^2 + 1$, some explanation should have been given as to how columns 2 and 3 were generated. But of course, the possibility exists that such an explanation was present originally. After all, it is apparent that the tablet is broken on the left-hand side. Perhaps it originally contained more columns of figures that might shed light on the entire tablet if we only had them. Here we enter upon immense possibilities, since the "vanished" portion of the tablet could have contained a huge variety of entries. To bring this open-ended discussion to a close, we look at what some experts in the area have to say.

In work that was apparently never published (see Buck, 1980, p. 344), D.L. Voils pointed out that tablets amounting to "teacher's manuals" have been found in which the following problem is set: *Find a number that yields a given number when its reciprocal is subtracted*. In modern terms this problem requires solving the equation

$$x - \frac{1}{x} = d,$$

where d is the given number. Obviously, if you were a teacher setting such a problem for a student, you would want the solution x to be such that both x and $1/x$ have terminating sexagesimal digits. So, if the solution is to be $x = p/q$, we

already see why we need both p and q to be products of 2, 3, and 5. This problem amounts to the quadratic equation $x^2 - dx - 1 = 0$, and its unique positive solution is $x = d/2 + \sqrt{1 + (d/2)^2}$. Column 1 of the tablet, which contains $(d/2)^2$ then appears as part of the solution process. It is necessary to take its square root and also the square root $\sqrt{1 + d^2/4}$ in order to find the solution $x = p/q$. This explanation seems to fit very well with the tablet. One could assume that the first column gives values of d that a teacher could use to set such a problem with the assurance that the pupil would get terminating sexagesimal expansions for both x and $1/x$. On the other hand, it does not fully explain why the tablet gives the numbers $p^2 - q^2$ and $p^2 + q^2$, rather than simply p and q , in subsequent columns. Doing our best for this theory, we note that columns 2 and 3 contain respectively the numerators of $x - 1/x$ and $x + 1/x$, and that their common denominator is the square root of the difference of the squares of these two numerators. Against that explanation is the fact that the Mesopotamians did not work with common fractions. The concepts of numerator and denominator to them would have been the concepts of dividend and divisor, and the final sexagesimal quotient would not display these numbers. The recipe for getting from columns 2 and 3 to column 1 would be first to square each of these columns, then find the reciprocal of the difference of the squares as a sexagesimal expansion, and finally, multiply the last result by the square in column 2.

In the course of a plea that historians look at Mesopotamian mathematics in its own terms rather than simply in relation to what came after, Robson (2001) examined several theories about the purpose of the tablet and gave some imaginative scenarios as to what may be in the lost portion of the tablet. Her conclusion, the only one justified by the present state of knowledge is that “the Mystery of the Cuneiform Tablet has not yet been fully solved.”²

And we have not claimed to solve it here. Plimpton 322 is a fascinating object of contemplation and serves as a *possible* example of an early interest in what we now call quadratic Diophantine equations. Without assuming that there is some continuous history between Plimpton 322 and modern number theory, we can still take quadratic Diophantine equations as a convenient starting point for discussing the history of number theory.

2. Ancient Greek number theory

Our knowledge of Pythagorean number theory is based on several sources, of which two important ones are Books 7–9 of Euclid’s *Elements* and a treatise on arithmetic by the neo-Pythagorean Nicomachus of Gerasa, who lived about 100 CE. Just as the *Sun Zi Suan Jing* preserves more of ancient Chinese arithmetic than the earlier *Jiu Zhang Suanshu*, it happens that the treatise of Nicomachus preserves more of Pythagorean lore than the earlier work of Euclid. For that reason, we discuss Nicomachus first.

The Pythagoreans knew how to find the greatest common divisor of two numbers. A very efficient procedure for doing so is described in Chapter 13 of Book 1 of Nicomachus’ *Arithmetica* and in Proposition 2 of Book 7 of Euclid’s *Elements*. This procedure, now known as the *Euclidean algorithm*, is what the Chinese called the *mutual-subtraction procedure*. Nicomachus applies it only to integers, any two

² In a posting at a mathematics history website, Robson noted that reciprocal pairs and cut-and-paste geometry seem to be the most plausible motives for the tablet.

of which naturally have 1 as a common divisor. Euclid, on the other hand, does not confine it to integers, but states the procedure for “magnitudes,” which may lack a common measure. It is significant that the procedure terminates if and only if there is a common measure, and Euclid makes use of that fact in discussing incommensurables. The algorithm was certainly invented long before the time of Euclid, however. Zverkina (2000) believes that this procedure could not have arisen intuitively, but must have come about as the result of solving specific problems, most likely the problem of reducing ratios by canceling a common divisor. It is used for that purpose in Chinese mathematics. What follows is a description of the general procedure.

For definiteness, we shall imagine that the two quantities whose greatest common *measure* is to be found are two lengths, say a and b . Suppose that a is longer than b . (If the two are equal, their common value is also their greatest common divisor.) The general procedure is to keep subtracting the smaller quantity from the larger until the remainder is equal to the smaller quantity or smaller than it. It is not difficult to show that the smaller quantity and the remainder have the same common measures as the smaller quantity and the larger. Hence one can start over with the smaller quantity and the remainder, which is no more than half of the larger quantity. Either this process terminates with an equal pair, or it continues and the pairs become arbitrarily small.

An example will make the procedure clear. Let us find the greatest common measure (divisor) of 26173996849 and 180569389. A common measure does exist: the integer 1. Since the repeated subtraction process amounts to division with remainder, we do it this way: $26173996849 \div 180569389$ is 144 with a remainder of 172004833. We then divide 180569389 by 172004833, getting a quotient of 1 and a remainder of 8564556. Next we divide 172004833 by 8564556, getting a quotient of 20 and a remainder of 713713. We then divide 8564556 by 713713 and get a quotient of 12 with no remainder, so that the greatest common divisor is 713713.

This computation can be arranged as follows:

$$\begin{array}{r}
 \begin{array}{cccc}
 12 & 20 & 1 & 144 \\
 713713 \overline{)8564556} & 172004833 \overline{)180569389} & 180569389 \overline{)26173996849} & \\
 \underline{8564556} & \underline{171291120} & \underline{172004833} & \underline{26001992016} \\
 0 & 713713 & 8564556 & 172004833
 \end{array}
 \end{array}$$

2.1. The *Arithmetica* of Nicomachus. In his first book Nicomachus makes the elementary distinction between odd and even numbers. Having made this distinction, he proceeds to refine it, distinguishing between even numbers divisible by 4 (evenly even) and those that are not (doubles of odd numbers). He goes on to classify odd numbers in a similar way, coming thereby to the concept of prime and composite numbers. Nicomachus also introduces what we now call pairs of *relatively prime numbers*. These are pairs of numbers that have no common prime divisor and hence no common divisor except 1. The notion of a relational property was difficult for Greek philosophers, and Nicomachus expresses the concept of relatively prime numbers in a confused manner, referring to three species of odd numbers: the prime and incomposite, the secondary and composite, and “the variety which, in itself is secondary and composite, but relatively is prime and incomposite.” This way of writing seems to imply that there are three kinds of integers, prime and incomposite, secondary and composite, and a third kind midway between the other two. It also seems to imply that one can look at an individual integer and classify it

into exactly one of these three classes. Such is not the case, however. The property of primeness is a property of a number alone. The property of being relatively prime is a property of a pair of numbers. On the other hand, the property of being relatively prime *to a given number* is a property of a number alone. Nicomachus explains the property in a rather wordy fashion in Chapter 13 of Book 1, where he gives a method of identifying prime numbers that has become famous as the *sieve of Eratosthenes*.

Nicomachus attributes this method to Eratosthenes. To use it, start with a list of all the odd numbers from 3 on, that is,

3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, . . .

From this list, remove every third number after 3, that is, remove 9, 15, 21, 27, 33, . . . These numbers are multiples of 3 and hence not prime. The reduced list is then

3, 5, 7, 11, 13, 17, 19, 23, 25, 29, 31, 35, 37, 41, 43, 47, 49, . . .

From this list, remove all multiples of 5 larger than 5. The first non-prime in the new list is $49 = 7 \cdot 7$. In this way, you can generate in short order a complete list of primes up to the square of the first prime whose multiples were not removed. Thus, after removing the multiples of 7, we have the list

3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61 . . .

The first non-prime in this list would be $11 \cdot 11 = 121$.

Nicomachus' point of view on this sieve was different from ours. Where we think of the *factors* of, say 60, as being 2, 2, 3, and 5, Nicomachus thought of the quotients by these numbers and products of them as the *parts* of a number. Thus, in his language 60 has the parts 30 (half of 60), 20 (one-third of 60), 15 (one-fourth of 60), 12 (one-fifth of 60), 10 (one-sixth of 60), 6 (one-tenth of 60), 5 (one-twelfth of 60), 4 (one-fifteenth of 60), 3 (one-twentieth of 60), 2 (one-thirtieth of 60) and 1 (one-sixtieth of 60). If these parts are added, the sum is 108, much larger than 60. Nicomachus called such a number *superabundant* and compared it to an animal having too many limbs. On the other hand, 14 is larger than the sum of its parts. Indeed, it has only the parts 7, 2, and 1, which total 10. Nicomachus called 14 a *deficient number* and compared it to an animal with missing limbs like the one-eyed Cyclops of the *Odyssey*. A number that is exactly equal to the sum of its parts, such as $6 = 1 + 2 + 3$, he called a *perfect number*. He gave a method of finding perfect numbers, which remains to this day the only way known to generate such numbers, although it has not been proved that there are no other such numbers. This procedure is also stated by Euclid: *If the sum of the numbers 1, 2, 4, . . . , 2^{n-1} is prime, then this sum multiplied by the last term will be perfect.* The modern statement of this fact is given in the exercises below. To see the recipe at work, start with 1, then double and add: $1 + 2 = 3$. Since 3 is prime, multiply it by the last term, that is, 2. The result is 6, a perfect number. Continuing, $1 + 2 + 4 = 7$, which is prime. Multiplying 7 by 4 yields 28, the next perfect number. Then, $1 + 2 + 4 + 8 + 16 = 31$, which is prime. Hence $31 \cdot 16 = 496$ is a perfect number. The next such number is $8128 = 64(1 + 2 + 4 + 8 + 16 + 32 + 64)$. In this way, Nicomachus was able to generate the first four perfect numbers. He seems to hint at a conjecture, but draws back from stating it explicitly:

When these have been discovered, 6 among the units and 28 in the tens, you must do the same to fashion the next. . . the result is 496, in the hundreds; and then comes 8,128 in the thousands, and so on, as far as it is convenient for one to follow [D'ooge, 1926, p. 211].³

This quotation seems to imply that Nicomachus expected to find one perfect number N_k having k digits. Actually, the fifth perfect number is 33,550,336, so we have jumped from four digits to eight here. The sixth is 8,589,869,056 (10 digits) and the seventh is 137,438,691,328 (12 digits), so that there is no regularity about the distribution of perfect numbers. Thus, Nicomachus was wise to refrain from making conjectures too explicitly. According to Dickson (1919, p. 8), later mathematicians, including Cardano, were less restrained, and this incorrect conjecture has been stated more than once.

For a topic that is devoid of applications, perfect numbers have attracted a great deal of attention from mathematicians. Dickson (1919) lists well over 100 mathematical papers devoted to this topic over the past few centuries. From the point of view of pure number theory, the main questions about them are the following: (1) Is there an odd perfect number?⁴ (2) Are all even perfect numbers given by the procedure described by Nicomachus?⁵ (3) Which numbers of the form $2^n - 1$ are prime? These are called *Mersenne primes*, after Marin Mersenne (1588–1648), who, according to Dickson (1919, pp. 12–13), first noted their importance, precisely in connection with perfect numbers. Obviously, n must itself be prime if $2^n - 1$ is to be prime, but this condition is not sufficient, since $2^{11} - 1 = 23 \cdot 89$. The set of known prime numbers is surprisingly small, considering that there are infinitely many to choose from, and the new ones being found tend to be Mersenne primes, mostly because that is where people are looking for them. The largest currently known prime (as of June 2004) is $2^{24036583} - 1$, only the forty-first Mersenne prime known.⁶ It was found on May 15, 2004 by the GIMPS (Great Internet Mersenne Prime Search) project, which links over 200,000 computers via the Internet and runs prime-searching software in the background of each while their owners are busy with their own work. This prime has 7,235,733 decimal digits. The fortieth Mersenne prime, $2^{20996011} - 1$, was found on November 17, 2003; it has 6,320,430 decimal digits. In contrast, the largest known non-Mersenne prime is $3 \cdot 2^{303093} + 1$, found by Jeff Young in 1998.⁷ It is rather tiny in comparison with the last few Mersenne primes discovered, having “only” 91,241 decimal digits.

Beginning in Chapter 6 of Book 2, Nicomachus studies figurate numbers: polygonal numbers through heptagonal numbers, and then polyhedral numbers. These numbers are connected with geometry, with an identification of the number 1 with a geometric point. To motivate this discussion Nicomachus speculated that the

³ D'ooge illustrates the procedure in a footnote, but states erroneously that 8191 is not a prime.

⁴ The answer is unknown at present.

⁵ The answer is yes. The result is amazingly easy to prove, but no one seems to have noticed it until a posthumous paper of Leonhard Euler gave a proof. Victor-Amédée Lebesgue (1791–1875) published a short proof in 1844.

⁶ The reader will correctly infer from previous footnotes that exactly 41 perfect numbers are now known.

⁷ See his article “Large primes and Fermat factors” in *Mathematics of Computation*, **67** (1998), 1735–1738, which gives a method of finding probable primes of the form $k \cdot 2^n + 1$.

simplest way to denote any integer would be repeating a symbol for 1 an appropriate number of times. Thus, he said, the number 5 could be denoted $\alpha\alpha\alpha\alpha\alpha$. This train of thought, if followed consistently, would lead back to a notation even more primitive than the hieroglyphic notation for numbers, since it would use only the symbol for units and discard the symbols for higher powers of 10. The Egyptians had gone beyond this principle in their hieratic notation, and the standard Greek notation was essentially a translation of the hieratic into the Greek alphabet. You can easily see where this speculation leads. The outcome is shown in Fig. 1, which illustrates triangular, square, pentagonal, and hexagonal numbers but using dots instead of the letter α . Observe that the figures are *not* associated with regular polygons except in the case of triangles and squares. The geometry alone makes it clear that a square number is the sum of the corresponding triangular number and its predecessor. Similarly, a pentagonal number is the sum of the corresponding square number and the preceding triangular number, a hexagonal number is the sum of the corresponding pentagonal number and the preceding triangular number, and so forth. This is the point at which modern mathematics parts company with Nicomachus, Proclus, and other philosophers who push analogies further than the facts will allow. As Nicomachus states at the beginning of Chapter 7:

The point, then, is the beginning of dimension, but not itself a dimension, and likewise the beginning of a line, but not itself a line; the line is the beginning of surface, but not surface; and the beginning of the two-dimensional, but not itself extended in two dimensions. . . Exactly the same in numbers, unity is the beginning of all number that advances unit by unit in one direction; linear number is the beginning of plane number, which spreads out like a plane in one more dimension. [D'ooge, 1926, p. 239]

This mystical mathematics was transmitted to Medieval Europe by Boethius. It is the same kind of analogical thinking found in Plato's *Timaeus*, where it is imagined that atoms of fire are tetrahedra, atoms of earth are cubes, and so forth. Since the Middle Ages, this topic has been of less interest to mathematicians. The phrase *of less interest*—rather than *of no interest*—is used advisedly here: There are a few theorems about figurate numbers in modern number theory, and they have some connections with analysis as well. For example, a formula of Euler asserts that

$$\prod_{k=1}^{\infty} (1 - x^k) = \sum_{n=-\infty}^{\infty} (-1)^n x^{n(3n-1)/2}.$$

Here the exponents on the right-hand side range over the pentagonal numbers for n positive. By making this formula the definition of the n th pentagonal number for negative n , we thereby gain an interesting formula that can be stated in terms of figurate numbers. Carl Gustav Jacobi (1804–1851) was pleased to offer a proof of this theorem as evidence of the usefulness of elliptic function theory. Even today, these numbers crop up in occasional articles in graph theory and elsewhere.

2.2. Euclid's number theory. Euclid devotes his three books on number theory to divisibility theory, spending most of the time on proportions among integers and on prime and composite numbers. Only at the end of Book 9 does he prove a theorem of a different sort, giving the method of searching for perfect numbers described above. It is interesting that Euclid does not mention figurate numbers. Although

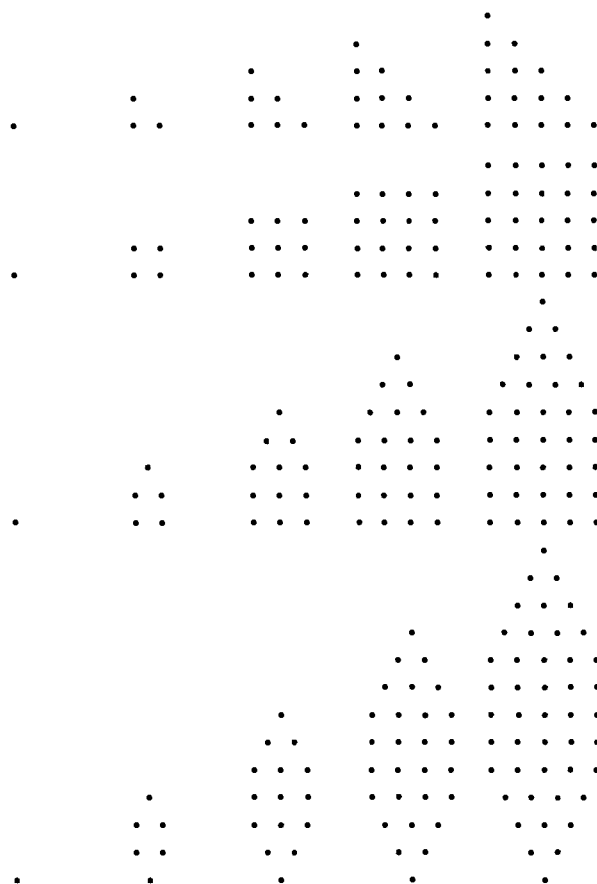


FIGURE 1. Figurate numbers. Top row: triangular numbers $T_n = n(n+1)/2$. Second row: square numbers $S_n = n^2$. Third row: pentagonal numbers $P_n = n(3n-1)/2$. Bottom row: hexagonal numbers $H_n = n(2n-1)$.

the Pythagorean and Platonic sources of Euclid's treatise are obvious, Euclid appears to the modern eye to be much more a mathematician than Pythagoras or Plato, much less addicted to flights of fanciful speculation on the nature of the universe. In fact, he never mentions the universe at all and suggests no practical applications of the theorems in his *Elements*.

Book 7 develops proportion for positive integers as part of a general discussion of how to reduce a ratio to lowest terms. The notion of relatively prime numbers is introduced, and the elementary theory of divisibility is developed as far as finding least common multiples and greatest common factors. Book 8 resumes the subject of proportion and extends it to squares and cubes of integers, including the interesting theorem that the mean proportional of two square integers is an integer (Proposition 11), and between any two cubes there are two such mean proportionals (Proposition 12): for example, $25 : 40 :: 40 : 64$, and $27 : 45 :: 45 : 75 :: 75 : 125$. Book 9 continues this topic; it also contains the famous theorem that there are infinitely many primes (Proposition 20) and ends by giving the only known method

of constructing perfect numbers (Proposition 36), quoted above. No perfect number has yet been found that is *not* generated by this procedure, although no proof exists that all perfect numbers are of this form. Any exception would have to be an odd number, since it is known (see Problem 7.8) that all even perfect numbers are of this form.

From the modern point of view, Euclid's number theory is missing an explicit statement of the *fundamental theorem of arithmetic*. This theorem, which asserts that every positive integer can be written in only one way as a product of prime numbers, can easily be deduced from Book 7, Proposition 24: *If two numbers are relatively prime to a third, their product is also relatively prime to it*. However, modern historians (Knorr, 1976) have pointed out that Euclid doesn't actually prove the fundamental theorem.

2.3. The *Arithmetica* of Diophantus. Two works of Diophantus have survived in part, a treatise on polygonal numbers and the work for which he is best known, the *Arithmetica*. Like many other ancient works, these two works of Diophantus survived because of the efforts of a ninth-century Byzantine mathematician named Leon, who organized a major effort to copy and preserve these works. There is little record of the influence the works of Diophantus may have exerted before this time.

According to the introduction to the *Arithmetica*, this work consisted originally of 13 books, but until recently only six were known to have survived; it was assumed that these were the first six books, on which Hypatia wrote a commentary. However, more books were recently found in an Arabic manuscript that the experts say is a translation made very early—probably in the ninth century. Sesiano (1982) stated that these books are in fact the books numbered 4 to 7, and that the books previously numbered 4 to 6 must come after them.

Diophantus begins with a small number of determinate problems that illustrate how to think algebraically, in terms of expressions involving a variable. Since these problems belong properly to algebra, they are discussed in Chapter 14. Indeterminate problems, which are number theory because the solutions are required to be rational numbers (the only kind recognized by Diophantus), begin in Book 2. A famous example of this type is Problem 8 of Book 2, *to separate a given square number into two squares*. Diophantus illustrates this problem using the number 16 as an example. His method of solving this problem is to express the two numbers in terms of a single unknown, which we shall denote ς , in such a way that one of the conditions is satisfied automatically. Thus, letting one of the two squares be ς^2 , which Diophantus wrote as Δ^v (as explained in Chapter 14), he noted that the other will automatically be $16 - \varsigma^2$. To get a determinate equation for ς , he assumes that the other number to be squared is 4 less than an unspecified multiple of ς . The number 4 is chosen because it is the square root of 16. In our terms, it leads to a quadratic equation one of whose roots is zero, so that the other root can be found by solving a linear equation. As we would write it, assuming that $16 - \varsigma^2 = (k\varsigma - 4)^2$, we find that $(k^2 + 1)\varsigma^2 = 8k\varsigma$, and—cancelling ς , since Diophantus does not operate with 0—we get $\varsigma = 8k/(k^2 + 1)$. This formula generates a whole infinite family of solutions of the equation that we would call $x^2 + y^2 = 16$ via the identity

$$\left(\frac{8k}{k^2 + 1}\right)^2 + \left(\frac{4(k^2 - 1)}{k^2 + 1}\right)^2 = 16.$$

You may be asking why it was necessary to use a square number (16) here. Why not separate any positive rational number, say 5, into a sum of two squares? If you look carefully at the solution, you will see that Diophantus had to make the constant term drop out of the quadratic equation, and that could only be done by introducing the square root of the given number.

Diophantus' procedure is slightly less general than what we have just shown, although his illustrations show that he knows the general procedure and could generate other solutions. In his illustration he assumes that the other square is $(2\varsigma - 4)^2$. Since this number must be $16 - \varsigma^2$, he finds that $4\varsigma^2 - 16\varsigma + 16 = 16 - \varsigma^2$, so that $\varsigma = \frac{16}{5}$. It is clear that this procedure can be applied very generally, showing an infinite number of ways of dividing a given square into two other squares.

At first sight it appears that number theory really is not involved in this problem, that it is a matter of pure algebra. However, the topic of the problem naturally leads to other questions that definitely do involve number theory, that is, the theory of divisibility of integers. The most obvious one is the problem of finding *all possible* representations of a positive rational number as the sum of the squares of two rational numbers. One could then generalize and ask how many ways a given rational number can be represented as the sum of the cubes or fourth powers, and so forth, of two rational numbers. Those of a more Pythagorean bent might ask how many ways a number can be represented as a sum of triangular, pentagonal, or hexagonal numbers. In fact, all of these questions have been asked, starting in the seventeenth century.

The problem just solved achieved lasting fame when Fermat, who was studying the *Arithmetica*, remarked that the analogous problem for cubes and higher powers had no solutions; that is, one cannot find positive integers x , y , and z satisfying $x^3 + y^3 = z^3$ or $x^4 + y^4 = z^4$, or, in general $x^n + y^n = z^n$ with $n > 2$. Fermat stated that he had found a proof of this fact, but unfortunately did not have room to write it in the margin of the book. Fermat never published any general proof of this fact, although the special case $n = 4$ is a consequence of a method of proof developed by Fermat, known as the method of infinite descent. The problem became generally known after 1670, when Fermat's son published an edition of Diophantus' work along with Fermat's notes. It was a tantalizing problem because of its comprehensibility. Anyone with a high-school education in mathematics can understand the statement of the problem, and probably the majority of mathematicians dreamed of solving it when they were young. Despite the efforts of hundreds of amateurs and prizes offered for the solution, no correct proof was found for more than 350 years. On June 23, 1993, the British mathematician Andrew Wiles announced at a conference held at Cambridge University that he had succeeded in proving a certain conjecture in algebraic geometry known as the Shimura-Taniyama conjecture, from which Fermat's conjecture is known to follow. This was the first claim of a proof by a reputable mathematician using a technique that is known to be feasible, and the result was tentatively endorsed by other mathematicians of high reputation. After several months of checking, some doubts arose. Wiles had claimed in his announcement that certain techniques involving what are called Euler systems could be extended in a particular way, and this extension proved to be doubtful. In collaboration with another British mathematician, Richard Taylor, Wiles eventually found an alternative approach that simplified the proof considerably, and there is now no doubt among the experts in number theory that the problem has been solved.

To give another illustration of the same method, we consider the problem following the one just discussed, that is, Problem 9 of Book II: *to separate a given number that is the sum of two squares into two other squares*. (That is, given one representation of a number as a sum of two squares, find a new representation of the same type.) Diophantus shows how to do this using the example $13 = 2^2 + 3^2$. He lets one of the two squares be $(\varsigma + 2)^2$ and the other $(2\varsigma - 3)^2$, resulting in the equation $5\varsigma^2 - 8\varsigma = 0$. Thus, $\varsigma = \frac{8}{5}$, and indeed $(\frac{18}{5})^2 + (\frac{1}{5})^2 = 13$. It is easy to see here that Diophantus is deliberately choosing a form for the solution that will cause the constant term to drop out. This amounts to a general method, used throughout the first two books, and based on the proportion

$$(a + Y) : X = X : (a - Y)$$

for solving the equation $X^2 + Y^2 = a^2$.

The method Diophantus used to solve such problems in his first two books was conjectured by Maximus Planudes (1255–1305) and has recently been explained in simple language by Christianidis (1998).

Some of Diophantus' indeterminate problems reach a high degree of complexity. For example, Problem 19 of Book 3 asks for four numbers such that if any of the numbers is added to or subtracted from the square of the sum of the numbers, the result is a square number. Diophantus gives the solutions as

$$\frac{17,136,600}{163,021,824}, \frac{12,675,000}{163,021,824}, \frac{15,615,600}{163,021,824}, \frac{8,517,600}{163,021,824}.$$

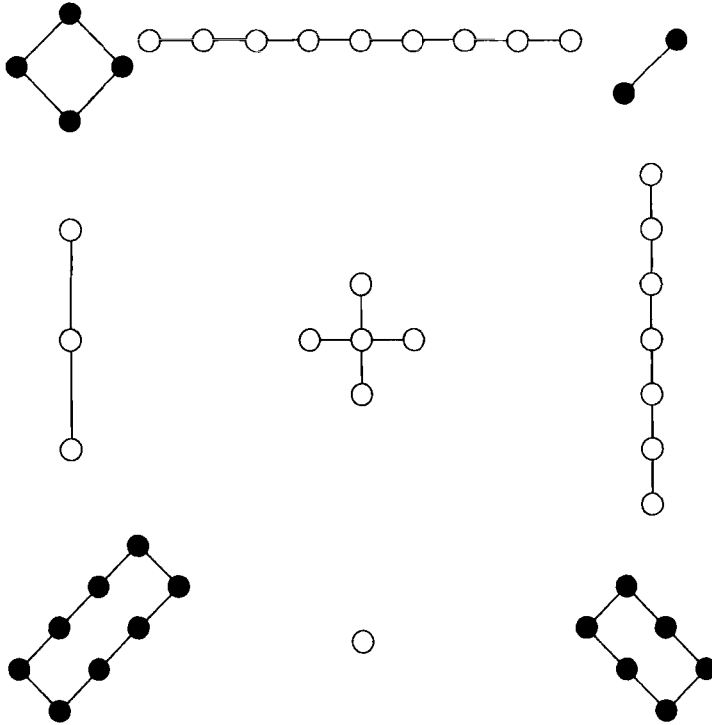
3. China

Although figurate numbers were not a topic of interest to early Chinese mathematicians, there was always in China a great interest in the use of numbers for divination. According to Li and Du (1987, pp. 95–97), the magic square

4	9	2
3	5	7
8	1	6

appears in the treatise *Shushu Ji* (*Memoir on Some Traditions of the Mathematical Art*) by the sixth-century mathematician Zhen Luan. In this figure each row, column, and diagonal totals 15. In the early tenth century, during the Song Dynasty, a connection was made between this magic square and a figure called the *Luo-chu-shu* (*book that came out of the River Lo*) found in the famous classic work *I Ching*, which was mentioned in connection with divination in Chapter 1. The *I Ching* states that a tortoise crawled out of the River Lo and delivered to the Emperor Yu the diagram in Fig. 2. Because of this connection, the diagram came to be called the *Luo-shu* (*Luo book*). Notice that the purely numerical aspects of the magic square are enhanced by representing the even (female, ying) numbers as solid disks and the odd (male, yang) numbers as open circles. Like so much of number theory, the theory of magic squares has continued to attract attention from specialists, all the while remaining essentially devoid of any applications. In this particular case, the interest has come from specialists in combinatorics, for whom magic squares and Latin squares form a topic of continuing research.

Another example of the use of numbers for divination comes from the last problem (Problem 36 of Chapter 3) of the *Sun Zi Suan Jing*. The data for the problem are very simple. A woman, aged 29, is pregnant. The period of human

FIGURE 2. The *Luo-shu*.

gestation is nine months. The problem is to determine the gender of the unborn child. In what is apparently an echo of the *I Ching* method of divination, the author begins with 49 (the number of yarrow stalks remaining after the first one has been laid down to begin the divination process). He then says to add the number of months of gestation, then subtract the woman's age. From the remainder (difference?) one is then to subtract successively 1 (heaven), 2 (Earth), 3 (man), 4 (seasons), 5 (phases), 6 (musical tones), 7 (stars in the Dipper), 8 (wind directions), and 9 (provinces of China under the Emperor Yu) and then use the final difference to determine the gender.⁸

The nature of divisibility for integers is also studied in Chinese treatises, in particular in the *Sun Zi Suan Jing*, which contains the essence of the result still known today as the Chinese remainder theorem. It was mentioned above that in general the *Sun Zi Suan Jing* is more elementary than the earlier *Jiu Zhang Suanshu*, but this bit of number theory is introduced for the first time in the *Sun Zi Suan Jing*. The problem asks for a number that leaves a remainder of 2 when divided by 3, a remainder of 3 when divided by 5, and a remainder of 2 when divided by 7. As in the case of Diophantus, the problem appears to be algebra, but it also involves the notion of divisibility with specified remainders. The assertion

⁸ Although no explanation is given in the translation by Lam and Ang (1992, p. 182), and no value is given for the final difference in this problem, the child is said by the author to be male. Perhaps the subtracting of successive integers was meant to continue only until the number left was smaller than the next number to be subtracted. In the present case, that number would be 1, resulting after 7 was subtracted. This interpretation seems to make sense; otherwise, the result of the procedure would be determined entirely by the parity of the woman's age.

that any number of such congruences can be solved simultaneously if the divisors are all pairwise relatively prime is the content of what we know now as the Chinese remainder theorem. According to Dickson, (1920, p. 57) this name arose when the mathematically astute British missionary Alexander Wylie (1815–1887) wrote an article on it in the English-language newspaper *North China Herald* in 1852. By that time the result was already known in Europe, having been discovered by Gauss and published in his *Disquisitiones arithmeticae* (Art. 36) in 1836.

Sun Zi's answer to this problem shows that he knew a general method of proceeding. He says, "Since the remainder on division by 3 is 2, take 140. The remainder on division by 5 is 3, so take 63. The remainder on division by 7 is 2, so take 30. Add these numbers, getting 233. From this subtract 210, getting the answer as 23."

It is just possible that the reader may not discern the underlying reasoning here, and so a bit of explanation may help. Sun Zi reasons that all multiples of $35 = 7 \cdot 5$ will leave a remainder of zero when divided by 5 or 7. He therefore took a multiple of 35, $140 = 35 \cdot 4$, that leaves a remainder of 2 when divided by 3. One may well ask why he didn't simply take 35 itself, since it also leaves a remainder of 2 when divided by 3. The seemingly clumsy use of 140 instead reveals still more of Sun Zi's thought processes. He must have looked first for a multiple of 35 that leaves a remainder of 1 when divided by 3, found it to be 70, then multiplied it by 2.⁹ Similarly, 63 is the smallest multiple of $21 = 7 \cdot 3$ that leaves a remainder of 3 when divided by 5, and 30 is the smallest multiple of $15 = 3 \cdot 5$ that leaves a remainder of 2 when divided by 7. Adding all three numbers, we get the number 233, which leaves the desired remainders on all three divisions. Sun Zi also knew that any multiple of $105 = 3 \cdot 5 \cdot 7$ could be added or subtracted without affecting any of the remainders. Hence subtracting 210 produced the smallest possible solution. It is obvious from this explanation that Sun Zi's method is perfectly general and can be used to find all possible solutions to such problems. What is concealed in his exposition is the general hypothesis that the divisors must be pairwise relatively prime. Sun Zi does not discuss this concept, but obviously he must have encountered cases where such problems cannot be solved. Almost certainly he would have traced the difficulty back to the existence of common factors among the divisors.

The importance of this kind of problem to the Chinese was not merely theoretical. Given that the ratio of a month—the time between two successive full moons—to a year is 19:235, questions involving calendars lead very often to finding numbers that leave a given remainder when divided by 19 and another given remainder when divided by 235. For example, suppose we know that the moon was full on June 1, 1996. What is the next year on which it will be full on June 4? (See Problem 7.10.)

The secret of solving problems of this sort is the Euclidean algorithm. This algorithm was known in China from the first century CE (see Shen, 1988) and used to solve a variety of problems, including the conversion of a long decimal expansion into a common fraction approximation with a small denominator (see Problem 7.12).

⁹ We can assume that Sun Zi found 70 by trial and error. The appropriate multiple would not be so easy if the divisors involved had seven digits each. A method for handling such harder cases was discovered by the Hindus and is discussed below.

4. India

The *Sulva Sūtras* contain rules for finding Pythagorean triples of integers, such as (3, 4, 5), (5, 12, 13), (8, 15, 17), and (12, 35, 37). It is not certain what practical use these arithmetic rules had. They may have been motivated by religious ritual. A Hindu home was required to have three fires burning at three different altars. The three altars were to be of different shapes, but all three were to have the same area. These conditions led to certain “Diophantine” problems, a particular case of which is the generation of Pythagorean triples, so as to make one square integer equal to the sum of two others.

One class of mathematical problems associated with altar building involves an altar of prescribed area in layers. In one problem from the *Bodhayana Sūtra* the altar is to have five layers of bricks, each layer containing 21 bricks. Now one cannot simply divide a pile of 105 identical bricks into five layers and pile them up. Such a structure would not be stable. It is necessary to stagger the edges of the bricks. So that the outside of the altar will not be jagged, it is necessary to have at least two different sizes of bricks. The problem is to decide how many different sizes of bricks will be needed and how to arrange them. Assuming an area of one square unit (actually, the unit is 1 square *vyayam*, about 7 square meters), the author suggests using three kinds of square bricks, of areas $\frac{1}{36}$, $\frac{1}{16}$, and $\frac{1}{9}$ square unit. The first, third, and fifth layers are to have nine of the first kind and 12 of the second. The second and fourth layers get 16 of the first kind and five of the third.

4.1. Varahamihira’s mystical square. According to Hayashi (1987), around the year 550 the mathematician Varahamihira wrote the *Brhatsamhita*, a large book devoted mainly to divination. However, Chapter 76 also discusses the mixing of perfumes from 16 substances, grouped in fours and mixed according to proportions given by the rows of the following square array:

2	3	5	8
5	8	2	3
4	1	7	6
7	6	4	1

Thus, the mysticism surrounding these squares penetrated even practical aspects of life. Hayashi notes that the Sanskrit word for the square itself, *kacchaputa*, means a box with compartments, but originally meant a tortoise shell. The resemblance to the *Luo Shu* is probably a coincidence.

4.2. Aryabhata I. In verses 32 and 33 of the *Aryabhatiya* we find a method of solving problems related to the problem of Sun Zi that leads to the Chinese remainder theorem. However, the context of the method and the description leave much to be desired in terms of clarity. It would have helped if Aryabhata had included specific examples. Such examples were provided by later commentators, and the process was described more clearly by Brahmagupta.

4.3. Brahmagupta. A century after Aryabhata, Brahmagupta called the method the *kuttaka* (pulverizer). We shall exclude certain complications in Brahmagupta’s presentation and present the method as simply as possible. The *kuttaka* provides the following visual implementation of an algorithm for solving the equation $ax = by + c$, with $b > a > 0$ and a and b relatively prime. As an example, we shall find all

solutions of the equation $4x = 23y + 5$. First, we carry out the Euclidean algorithm until 1 appears as a remainder:

$$\begin{aligned} 23 &= 5 \cdot 4 + 3, \\ 4 &= 1 \cdot 3 + 1. \end{aligned}$$

We then write the quotients (5 and 1 in this case) from the Euclidean algorithm in a column, and beneath them we write the additive term c if the number of quotients is even (in this case, two), otherwise $-c$. At the bottom of the column we write a 0. This zero is inserted so that the same transformation rule applies at the beginning as in all other steps of the algorithm. We then reduce the number of rows successively by operating on the bottom three rows at each stage. The second-from-last row is replaced by its product with the next-to-last row plus the last row; the next-to-last row is simply copied, and the last row is discarded. Thus to solve this system the *kuttaka* method amounts to the transformations

$$\begin{array}{rcccl} 5 & & & & \\ 1 & & 5 & & \\ 5 & \longrightarrow & 5 & \longrightarrow & 30 \\ 0 & & 5 & & 5 \end{array}.$$

This column now gives x and y , and indeed, $4 \cdot 30 = 23 \cdot 5 + 5$. Diophantus showed how to find a particular solution of such a congruence. Brahmagupta, however, found *all* the solutions. He took the solutions x and y obtained by the *kuttaka* method, which were generally quite large numbers, divided x by b and y by a , replaced them by the remainders, and gave the general x and y as a pair of arithmetic sequences with differences b and a , respectively. In the present case, the general solution is $x = 30 + 23k$, $y = 5 + 4k$. The smallest positive solution $x = 7$, $y = 1$, is obtained by taking $k = -1$.

Brahmagupta's rule for finding the solutions is more complicated than the discussion just given, since he does not assume that the numbers a and b are relatively prime. However, the greater generality is only apparent. If the greatest common divisor of a and b is not a factor of c , the problem has no solution; if it is a factor of c , it can be divided out of the problem.

Brahmagupta also considers such equations with negative data and is not in the least troubled by this complication. It seems clear that the name *pulverizer* was applied because the original data are repeatedly broken down by the Euclidean algorithm (they are "pulverized").

Astronomical applications. It was mentioned above that this kind of remainder arithmetic, which we now call the *theory of linear congruences*, has applications to the calendar. Brahmagupta proposed the problem of finding the (integer) number of elapsed days when Jupiter is $22^\circ, 30'$ into the sign of Aries¹⁰ (Colebrooke, 1817, p. 334). Brahmagupta converted the $22^\circ, 30'$ into $1350'$. He had earlier taken the sidereal period of Saturn to be 30 years and to be a common multiple of all cycles.

¹⁰ Obviously, Jupiter will pass this point once in each revolution, but it will reach *exactly* this point at the expiration of an *exact* number of days (no fractional hours or minutes) only once in a *yuga*, which is a common period for all the heavenly bodies. Brahmagupta took 30 years as a *yuga*, but his method is general and will yield better results if a more accurate *yuga* is provided by observation. He says that the value of 30 years is given for a *yuga* only to make the computation easier.

On that basis Jupiter was given (inaccurately) a sidereal period of 10 years. Again inaccurately, using a year of $365\frac{1}{3}$ days, we find that Jupiter undergoes three cycles in 10,960 days. Thus, in one day, Jupiter moves $3/10960$ of a revolution. Since there are $360 \cdot 60 = 21600$ minutes in a revolution, we find that each day Jupiter moves $64800/10960 = 810/137$ minutes. The problem, then, is to solve $810x/137 = 21,600y + 1350$ or dividing out the common factor of 270, $3x = 137 \cdot (80y + 5)$; that is, $3x = 10,960y + 685$. Here x will be the number of days elapsed in the cycle and y the number of revolutions that Jupiter will have made. The Euclidean algorithm yields $10960 = 3 \cdot 3653 + 1$, so that we have

$$\begin{array}{rcl} 3653 & & \\ -685 & \longrightarrow & -2502305 \\ 0 & & -685 \end{array}$$

That is, $x = -2502305 + 10960t$ and $y = -685 + 3t$. The smallest positive solution occurs when $t = 229$: $x = 7535$, $y = 2$, which is Brahmagupta's answer.

Brahmagupta illustrated the formulas for right triangles by creating Pythagorean triples. In Chapter 12 of the *Brahmasphutasiddhanta* (Colebrooke, 1817, p. 306) he gives the rule that "the sum of the squares of two unlike quantities are the sides of an isosceles triangle; twice the product of the same two quantities is the perpendicular; and twice the difference of their square is the base." This rule amounts to the formula $(a^2 + b^2)^2 = (2ab)^2 + (a^2 - b^2)^2$, but it is stated as if the right triangle has been doubled by gluing another copy to the side of length $a^2 - b^2$, thereby producing an isosceles triangle. The relation stated is a purely geometric relation, showing (in our terms) that the sides and altitude of an isosceles triangle of any shape can be generated by choosing the two lengths a and b suitably.

Brahmagupta also considered generalizations of the problem of Pythagorean triples to a more general equation called¹¹ *Pell's equation* and written $y^2 - Dx^2 = 1$. He gives a recipe for generating a new equation of this form and its solutions from a given solution. The recipe proceeds by starting with two rows of three entries, which we shall illustrate for the case $D = 8$, which has the solution $x = 1$, $y = 3$. We write

$$\begin{array}{ccc} 1 & 3 & 1 \\ 1 & 3 & 1 \end{array}$$

The first column contains x , called the *lesser solution*, the second contains y , called the *greater solution*, and the third column contains the additive term 1. From these two rows a new row is created whose first entry is the sum of the cross-multiplied first two columns, that is $1 \cdot 3 + 3 \cdot 1 = 6$. The second entry is the product of the second entries plus 8 times the product of the first entries, that is $3 \cdot 3 + 8 \cdot 1 \cdot 1 = 17$, and the third row is the product of the third entries. Hence we get a new row 6 17 1, and indeed $8 \cdot 6^2 + 1 = 289 = 17^2$. In our terms, this says that if $8x^2 + 1 = y^2$ and $8u^2 + 1 = v^2$, then $8(xv + yu)^2 + 1 = (8xu + yv)^2$. More generally, Brahmagupta's rule says that if $ax^2 + d = y^2$ and $au^2 + c = v^2$, then

$$(1) \quad a(xv + yu)^2 + cd = (axu + yv)^2.$$

¹¹ Erroneously so-called, according to Dickson (1920, p. 341), who asserts that Fermat had studied the equation earlier than John Pell (1611–1685). However, the website at St. Andrew's University gives evidence that Euler's attribution of this equation to Pell was accurate. In any case, everybody agrees that the solutions of the equation were worked out by Lagrange, not Pell.

Although it is trivial to verify that this rule is correct using modern algebraic notation, one would like to know how it was discovered.¹² Although the route by which this discovery was made is not known, the *motivation* for studying the equation can be plausibly ascribed to a desire to approximate irrational square roots with rational numbers. Brahmagupta's rule with $c = d = 1$ gives a way of generating larger and larger solutions of the *same* Diophantine equation $ax^2 + 1 = y^2$. If you have two solutions (x, y) and (u, v) of this equation, which need not be different, then you have two approximations y/x and v/u for \sqrt{a} whose squares are respectively $1/x^2$ and $1/u^2$ larger than a . The new solution generated will have a square that is only $1/(xv + yu)^2$ larger than a . This aspect of the problem of Pell's equation turns out to have a close connection with its complete solution in the eighteenth century.

4.4. Bhaskara II. In his treatise *Vija Ganita (Algebra)*, Bhaskara states the rule for *kuttaka* more clearly than either Aryabhata or Brahmagupta had done and illustrates it with specific cases. For example, in Chapter 2 (Colebrooke, 1817, p. 162) he asks, "What is that multiplier which, when it is multiplied by 221 and 65 is added to the product, yields a multiple of 195?" In other words, solve the equation $195x = 221y + 65$. Dividing out 13, which is a common factor, reduces this equation to $15x = 17y + 5$. The *kuttaka*, whose steps are shown explicitly, yields as a solution $x = 40, y = 35$.

In his writing on algebra, Bhaskara considered many Diophantine equations. For example, in Section 4 of Chapter 3 of the *Lilavati* (Colebrooke, 1817, p. 27), he posed the problem of finding pairs of (rational) numbers such that the sum and difference of their squares are each 1 larger than a square. It would be interesting to know how he found the answer to this difficult problem. All he says is that the smaller number should be obtained by starting with any number, squaring, multiplying by 8, subtracting 1, then dividing by 2 and by the original number. The larger number is then obtained by squaring the smaller one, dividing by 2, and adding 1. In our terms, these recipes say that if u is any rational number, then

$$\left(8u^2 - 1 + \frac{1}{8u^2}\right)^2 \pm \left(4u - \frac{1}{2u}\right)^2 - 1$$

is the square of a rational number. The reader can easily verify that it is $(8u^2 - 1/(8u^2))^2$ when the positive sign is chosen and $(8u^2 - 2 + \frac{1}{8u^2})^2$ when the negative sign is taken.

Chapter 4 of the *Vija Ganita* contains many algebraic problems involved with solving triangles, interspersed with some pure Diophantine equations. One of the most remarkable (Colebrooke, 1817, p. 200) is the problem of finding four unequal (rational) numbers whose sum equals the sum of their squares or the sum of the cubes of which equals the sum of their squares. In the first case he gives $\frac{1}{3}, \frac{2}{3}, \frac{3}{3}, \frac{4}{3}$. In the second case he gives $\frac{3}{10}, \frac{6}{10}, \frac{9}{10}, \frac{12}{10}$. In both cases the numbers are in the proportion 1 : 2 : 3 : 4. These three extra conditions (three ratios of numbers) were deliberately added by Bhaskara so that the problem would become a determinate one.

The characteristic that makes problems like the preceding one easy is that the requirement imposed on the four numbers amounts to a single equation with

¹² Weil (1984, pp. 17, 83, 204) refers to Eq. 1 and the more general relation $(x^2 + Ny^2)(z^2 + Nt^2) = (xz \pm Nyt)^2 + N(xt \mp yz)^2$ as "Brahmagupta's identity" (his quotation marks).

more than one unknown. But Bhaskara also asks harder questions. For example (Colebrooke, 1817, p. 202): *Find two (rational) numbers such that the sum of the cubes is a square and the sum of the squares is a cube.* Bhaskara manages to find a solution using the trick of assigning the ratio a of the two numbers. It is necessary for the technique that this ratio satisfy $1 + a^3 = b^2$. Bhaskara chooses $a = 2$, $b = 3$.¹³ The smaller number is then chosen to be of the form $(1 + a^2)w^3$ for some number w . The sum of the squares will then be $(1 + a^2)^2w^6 + a^2(1 + a^2)^2w^6 = (1 + a^2)^3w^6 = ((1 + a^2)w^2)^3$, and the sum of their cubes will be $(1 + a^2)^3w^9 + a^3(1 + a^2)^3w^9 = (1 + a^2)^3b^2w^9 = b^2((1 + a^2)w^3)^3$. Hence, if w is chosen so that $(1 + a^2)w$ is a square, this will be a perfect square. The simplest choice obviously is $w = 1 + a^2$. In Bhaskara's example, that choice gives the pair 625 and 1250.

5. The Muslims

The Muslims continued the work of Diophantus in number theory. Abu-Kamil (ca. 850–ca. 930) wrote a book on “indeterminate problems” in which he studied quadratic Diophantine equations and systems of such equations in two variables. The first 38 problems that he studied are arranged in a very strict ordering of coefficients, exponents, and signs, making it a very systematic exposition of these equations. Later scholars noted the astonishing fact that the first 25 of these equations are what are now known as algebraic curves of genus 0, while the last 13 are of genus 1, even though the concept of genus of an algebraic curve is a nineteenth-century invention (Baigozhina, 1995).

Muslim mathematicians also went beyond what is in Euclid and Nicomachus, generalizing perfect numbers. In a number of articles, Rashed (see, for example, 1989) points out that a large amount of theory of abundant, deficient, and perfect numbers was assembled in the ninth century by Thabit ibn-Qurra and others, and that ibn al-Haytham (965–1040) was the first to state and attempt to prove that Euclid's formula gives all the even perfect numbers. Thabit ibn Qurra made an interesting contribution to the theory of amicable numbers. A pair of numbers is said to be *amicable* if each is the sum of the proper divisors of the other. The smallest such pair of numbers is 220 and 284. Although these numbers are not discussed by Euclid or Nicomachus, the commentator Iamblichus (see Dickson, 1919, p. 38) ascribed this notion to Pythagoras, who is reported as saying, “A friend is another self.” This definition of a friend is given by Aristotle in his *Nicomachean Ethics* (Bekker 1170b, line 7).

We mentioned above the standard way of generating perfect numbers, namely the Euclidean formula $2^{n-1}(2^n - 1)$, whenever $2^n - 1$ is a prime. Thabit ibn-Qurra found a similar way of generating pairs of amicable numbers. His formula is

$$2^n(3 \cdot 2^n - 1)(3 \cdot 2^{n-1} - 1) \text{ and } 2^n(9 \cdot 2^{2n-1} - 1),$$

whenever $3 \cdot 2^n - 1$, $3 \cdot 2^{n-1} - 1$, and $9 \cdot 2^{2n-1} - 1$ are all prime. The case $n = 2$ gives the pair 220 and 284. Whatever one may think about the impracticality of amicable numbers, there is no denying that Thabit's discovery indicates very

¹³ It was conjectured in 1844 by the Belgian mathematician Eugène Charles Catalan (1814–1894) that the only nonzero solutions to the Diophantine equation $p^m - q^n = 1$ are $q = 2 = n$, $p = 3 = n$. This conjecture was proved in 2002 by Predhu Mihailescu, a young mathematician at the Institute for Scientific Computing in Zürich. Lucky Bhaskara! He found the only possible solution.

profound insight into the divisibility properties of numbers. It is very difficult to imagine how he could have discovered this result. A conjecture, which cannot be summarized in a few lines, can be found in the article by Brentjes and Hogendijk (1989).

It is not clear how many new cases can be generated from this formula, but there definitely are some. For example, when $n = 4$, we obtain the amicable pair $17,296 = 16 \cdot 23 \cdot 47$ and $18,416 = 16 \cdot 1151$. Hogendijk (1985) gives Thabit ibn-Qurra's proof of his criterion for amicable numbers and points out that the case $n = 7$ generates the pair 9,363,584 and 9,437,056, which first appeared in Arabic texts of the fourteenth century.

Unlike some other number-theory problems such as the Chinese remainder theorem, which arose in a genuinely practical context, the theory of amicable numbers is an offshoot of the theory of perfect numbers, which was already a completely "useless" topic from the beginning. It did not seem useless to the people who developed it, however. According to M. Cantor (1880, p. 631) the tenth-century mystic al-Majriti recommended as a love potion writing the numbers on two sheets of paper and eating the number 284, while causing the beloved to eat the number 220. He claimed to have verified the effectiveness of this charm by personal experience! Dickson (1919, p. 39) mentions the Jewish scholar Abraham Azulai (1570–1643), who described a work purportedly by the ninth-century commentator Rau Nachshon, in which the gift of 220 sheep and 220 goats that Jacob sent to his brother Esau as a peace offering (Genesis 32:14) is connected with the concept of amicable numbers.¹⁴ In any case, although their theory seems more complicated, amicable numbers are easier to find than perfect numbers. Euler alone found 62 pairs of them (see Erdős and Dudley, 1983).

Another advance on the Greeks can be found in the work of Kamal al-Din al-Farisi, a Persian mathematician who died around 1320. According to Ağargün and Fletcher (1994), he wrote the treatise *Memorandum for Friends Explaining the Proof of Amicability*, whose purpose was to give a new proof of Thabit ibn-Qurra's theorem. Proposition 1 in this work asserts the existence (but not uniqueness) of a prime decomposition for every number. Propositions 4 and 5 assert that this decomposition is unique, that two distinct products of primes cannot be equal.

6. Japan

In 1627 Yoshida Koyu wrote a textbook of arithmetic called the *Jinkō-ki* (*Treatise on Large and Small Numbers*). This book contained a statement of what is known in modern mathematics as the *Josephus problem*. The Japanese version of the problem involves a family of 30 children choosing one of the children to inherit the parents' property. The children are arranged in a circle and count off by tens; the unlucky children who get the number 10 are eliminated; that is, numbers 10, 20, and 30 drop out. The remaining 27 children then count off again. The children originally numbered 11 and 22 will be eliminated in this round, and when the second round of numbering is complete, the child who was first will have the number 8. Hence the children originally numbered 3, 15, and 27 will be eliminated on the next

¹⁴ The peace offering was necessary because Jacob had tricked Esau out of his inheritance. But if the gift was symbolic and associated with amicable numbers, the story seems to imply that Esau was obligated to give Jacob 284 sheep and 284 goats. Perhaps there was an ulterior motive in the gift!

round, and the first child will start the following round as number 3. The problem is to see which child will be the last one remaining. Obviously, solving this problem in advance could be very profitable, as the original Josephus story indicates.¹⁵ The Japanese problem is made more interesting and more complicated by considering that half of the children belong to the couple and half are the husband's children by a former marriage. The wife naturally wishes one of her own children to inherit, and she persuades the husband to count in different ways on different rounds. The problem was reprinted by several later Japanese mathematicians.

The eighteenth-century mathematician Matsunaga Ryohitsu (1718–1749) discussed a variety of equations similar to the Pell equation and representations of numbers in general as sums and differences of powers. For example, his recipe for solving the equation $x^3 - y^3 = z^4$ was to take $z = m^3 - n^3$, then let $x = mz$, $y = nz$. But he also tackled some much more sophisticated problems, such as the problem of representing a given integer k as a sum of two squares and finding an integer that is simultaneously of the forms $y_1^2 + 69y_1 + 15$ and $y_2^2 + 72y_2 + 7$. Matsunaga gave the solution as 11,707, obtained by taking $y_1 = 79$, $y_2 = 78$.

7. Medieval Europe

In his *Liber quadratorum* (*Book of Squares*) Leonardo of Pisa (Fibonacci, 1170–1250) speculated on the difference between square and nonsquare numbers. In the prologue, addressed to the Emperor Frederick II, Leonardo says that he had been inspired to write the book because a certain John of Palermo, whom he had met at Frederick's court, had challenged him to find a square number such that if 5 is added to it or subtracted from it, the result is again a square. This question inspired him to reflect on the difference between square and nonsquare numbers. He then notes his pleasure on learning that Frederick had actually read one of his previous books and uses that fact as justification for writing on the challenge problem.

The *Liber quadratorum* is written in the spirit of Diophantus and shows a keen appreciation of the conditions under which a rational number is a square. Indeed, the ninth of its 24 propositions is a problem of Diophantus: *Given a nonsquare number that is the sum of two squares, find a second pair of squares having this number as their sum*. As mentioned above, this problem is Problem 9 of Book 2 of Diophantus.

The securest basis of Leonardo's fame is a single problem from his *Liber abaci*, written in 1202:

How many pairs of rabbits can be bred from one pair in one year given that each pair begins to breed in the second month after its birth, producing one new pair per month?

¹⁵ Josephus tells us that, faced with capture by the Romans after the fall of Jotapata, he and his Jewish comrades decided to commit mass suicide rather than surrender. Later commentators claimed that they stood in a circle and counted by threes, agreeing that every third soldier would be killed by the person on his left. The last one standing was duty bound to fall on his sword. According to this folk legend, Josephus immediately computed where he should position himself in order to be that last person, but decided to surrender instead of carrying out the bargain. Josephus himself, however, writes in *The Jewish Wars*, Book III, Chapter 8 that the order of execution was determined by drawing lots and that he and his best friend survived either by chance or by divine intervention in these lots. The mathematical problem we are discussing is also said to have been invented by Abraham ben Meir ibn Ezra (1092–1167), better known as Rabbi Ben Ezra, one of many Jewish scholars who flourished in the Caliphate of Cordoba.

By brute-force enumeration of cases, the author concludes that there will be 377 pairs, and “in this way you can do it for the case of infinite numbers of months.”

The sequence generated here (1, 1, 2, 3, 5, 8, ...), in which each term after the second is the sum of its two predecessors, has been known as the *Fibonacci sequence* ever since the *Liber abaci* was first printed in the nineteenth century. The Fibonacci sequence has been an inexhaustible source of identities. Many curious representations of its terms have been obtained, and there is a mathematical journal, the *Fibonacci Quarterly*, named in its honor and devoted to its lore. The Fibonacci sequence has been a rich source of interesting pure mathematics, but it has also had some illuminating practical applications, one of which is discussed in Problems 2.4–2.6.

Questions and problems

7.1. Compute the sexagesimal representation of the number

$$\left(\frac{p/q - q/p}{2}\right)^2$$

for the following pairs of integers (p, q) : (12, 5), (64, 27), (75, 32), (125, 54), and (9, 5). Then correct column 1 of Plimpton 322 accordingly.

7.2. On the surface the Euclidean algorithm looks easy to use, and indeed it is easy to use when applied to integers. The difficulty arises when it is applied to continuous objects (lengths, areas, volumes, weights). In order to execute a loop of this algorithm, you must be able to decide which element of the pair (a, b) is larger. But all judgments as to relative size run into the same difficulty that we encounter with calibrated measuring instruments: limited precision. There is a point at which one simply cannot say with certainty that the two quantities are either equal or unequal. Does this limitation have any practical significance? What is its theoretical significance? Show how it could give a wrong value for the greatest common measure even when the greatest common measure exists. How could it ever show that two quantities have *no* common measure?

7.3. The remainders in the Euclidean algorithm play an essential role in finding the greatest common divisor. The greatest common divisor of 488 and 24 is 8, so that the fraction 24/488 can be reduced to 3/61. The Euclidean algorithm generates two *quotients*, 20 and 3 (in order of generation). What is their relation to the two numbers? Observe the relation

$$\frac{1}{20 + \frac{1}{3}} = \frac{3}{61}$$

If you find the greatest common divisor of 23 and 56 (which is 1) this way, you will generate the quotients 2, 2, 3, 3. Verify that

$$\frac{23}{56} = \frac{1}{2 + \frac{1}{2 + \frac{1}{3 + \frac{1}{3}}}}$$

This expression is called the *continued fraction representation* representation of 23/56. Formulate a general rule for finding the continued fraction representation of a proper fraction.

7.4. Draw dot figures for the first five heptagonal and octagonal numbers. What kind of figure would you need for nonagonal numbers?

7.5. Prove the formulas given in the caption of Fig. 1 for T_n , S_n , P_n , and H_n . Then prove that $S_n = T_n + T_{n-1}$, $P_n = S_n + T_{n-1} = T_n + 2T_{n-1}$, $H_n = P_n + T_{n-1} = T_n + 3T_{n-1}$. If $P_{k,n}$ is the n th k -gonal number, give a general formula for $P_{k,n}$ in terms of k and n .

7.6. Prove that the Pythagorean procedure always produces a perfect number. That is, if $p = 2^n - 1$ is prime, then $N = 2^{n-1}p$ is perfect. This theorem is not difficult to prove nowadays, since the “parts” (proper divisors) of N are easy to list and sum.

7.7. Let N_n be the n th perfect number, so that $N_1 = 6$, $N_2 = 28$, $N_3 = 496$, $N_4 = 8128$. Assuming that all perfect numbers are given by the Pythagorean formula, that is, they are of the form $2^{n-1}(2^n - 1)$ when $2^n - 1$ is a prime, prove that $N_{n+1} > 16N_n$ if $n > 1$. Conclude that there cannot be more than one k -digit perfect number for each k .

7.8. (*V. A. Lebesgue's proof of Euler's theorem on even perfect numbers*) Suppose that the perfect number N has the prime factorization $N = 2^\alpha p_1^{n_1} \cdots p_k^{n_k}$, where p_1, \dots, p_k are distinct odd primes and α, n_1, \dots, n_k are nonnegative integers. Since N is perfect, the sum of *all* its divisors is $2N$. This means that

$$\begin{aligned} 2^{\alpha+1} p_1^{n_1} \cdots p_k^{n_k} &= (1 + 2 + \cdots + 2^\alpha)(1 + p_1 + \cdots + p_1^{n_1}) \cdots (1 + p_k + \cdots + p_k^{n_k}) \\ &= (2^{\alpha+1} - 1)(1 + p_1 + \cdots + p_1^{n_1}) \cdots (1 + p_k + \cdots + p_k^{n_k}). \end{aligned}$$

Rewrite this equation as follows:

$$\begin{aligned} (2^{\alpha+1} - 1)p_1^{n_1} \cdots p_k^{n_k} + p_1^{n_1} \cdots p_k^{n_k} &= \\ &= (2^{\alpha+1} - 1)(1 + p_1 + \cdots + p_1^{n_1}) \cdots (1 + p_k + \cdots + p_k^{n_k}), \\ p_1^{n_1} \cdots p_k^{n_k} + \frac{p_1^{n_1} \cdots p_k^{n_k}}{2^{\alpha+1} - 1} &= (1 + p_1 + \cdots + p_1^{n_1}) \cdots (1 + p_k + \cdots + p_k^{n_k}). \end{aligned}$$

Since the second term on the left must be an integer, it follows that $2^{\alpha+1} - 1$ must divide $p_1^{n_1} \cdots p_k^{n_k}$. This is not a significant statement if $\alpha = 0$ (N is an odd number). But if N is even, so that $\alpha > 0$, it implies that $2^{\alpha+1} - 1 = p_1^{m_1} \cdots p_k^{m_k}$ for integers $m_1 \leq n_1, \dots, m_k \leq n_k$, *not all zero*. Thus, the left-hand side consists of the two *distinct* terms $p_1^{n_1} \cdots p_k^{n_k} + p_1^{r_1} \cdots p_k^{r_k}$. It follows that the right-hand side must also be equal to this sum. Now it is obvious that the right-hand side *contains* these two terms. That means the sum of the remaining terms on the right-hand side must be zero. But since the coefficients of all these terms are positive, there *can be* only two terms on the right. Since the right-hand side obviously contains $(n_1 + 1)(n_2 + 1) \cdots (n_k + 1)$ terms, we get the equation

$$2 = (n_1 + 1)(n_2 + 1) \cdots (n_k + 1).$$

Deduce from this equation that N must be of the form $2^{n-1}(2^n - 1)$ and that $2^n - 1$ is prime.

7.9. Generalize Diophantus' solution to the problem of finding a second representation of a number as the sum of two squares, using his example of $13 = 2^2 + 3^2$ and letting one of the numbers be $(\zeta + 3)^2$ and the other $(k\zeta - 2)^2$.

7.10. Take as a unit of time $T = \frac{1}{235}$ of a year, about 37 hours, 18 minutes, say a day and a half in close approximation. Then one average lunar month is $M = 19T$, and one average solar year is $Y = 235T$. Given that the Moon was full on June 1, 1996, what is the next year in which it will be full on June 4? Observe that June 4 in whatever year that is will be 3 days ($2T$) plus an integer number of years. We are seeking integer numbers of months (x) and years (y), counting from June 1, 1996, such that $Mx = Yy + 2T$, that is (canceling T), $19x = 235y + 2$. Use the *kuttaka* to solve this problem and check your answer against an almanac. If you use this technique to answer this kind of question, you will get the correct answer most of the time. When the answer is wrong, it will be found that the full moon in the predicted year is a day earlier or a day later than the prescribed date. The occasional discrepancies occur because (1) the relation $M = 19T$ is not precise, (2) full moons occur at different times of day, and (3) the greatest-integer function is not continuous.

7.11. Use Bhaskara's method to find two integers such that the square of their sum plus the cube of their sum equals twice the sum of their cubes. (This is a problem from Chapter 7 of the *Vija Ganita*.)

7.12. The Chinese mutual-subtraction algorithm (the Euclidean algorithm) can be used to convert a decimal expansion to a common fraction and to provide approximations to it with small denominators. Consider, for example, the number $e \approx 2.71828$. By the Euclidean algorithm, we get

$$\begin{aligned}
 271,828 &= 2 \cdot 100,000 + 71,828 \\
 100,000 &= 1 \cdot 71,828 + 28,172 \\
 71,828 &= 2 \cdot 28,172 + 15,484 \\
 28,172 &= 1 \cdot 15,484 + 12,688 \\
 15,484 &= 1 \cdot 12,688 + 2,796 \\
 12,688 &= 4 \cdot 2,796 + 1,504 \\
 2,796 &= 1 \cdot 1,504 + 1,292 \\
 1,504 &= 1 \cdot 1,292 + 212 \\
 1,292 &= 6 \cdot 212 + 20 \\
 212 &= 10 \cdot 20 + 12 \\
 20 &= 1 \cdot 12 + 8 \\
 12 &= 2 \cdot 8 + 4 \\
 8 &= 2 \cdot 4
 \end{aligned}$$

Thus the greatest common divisor of 271,828 and 100,000 is 4, and if it is divided out of all of these equations, the quotients remain the same. We can thus write

$$2.71828 = \frac{271828}{100000} = \frac{67957}{25000} = 2 + \frac{1}{1 + \frac{1}{2 + \frac{1}{1 + \frac{1}{1 + \frac{1}{4 + \dots}}}}}$$

The first few partial fractions here give

$$\begin{aligned}
 2 + \frac{1}{1} &= 3, \\
 2 + \frac{1}{1 + \frac{1}{2}} &= 2\frac{2}{3} = \frac{8}{3} = 2.666\dots, \\
 2 + \frac{1}{1 + \frac{1}{2 + \frac{1}{1}}} &= \frac{11}{4} = 2.75, \\
 2 + \frac{1}{1 + \frac{1}{2 + \frac{1}{1 + \frac{1}{1}}}} &= 2\frac{5}{7} = \frac{19}{7} = 2.714285712485\dots, \\
 2 + \frac{1}{1 + \frac{1}{2 + \frac{1}{1 + \frac{1}{1 + \frac{1}{4}}}}} &= 2 + \frac{23}{32} = \frac{87}{32} = 2.71875,
 \end{aligned}$$

so that the approximations get better and better. Do the same with $\pi \approx 3.14159265$, and calculate the first five approximate fractions. Do you recognize any of these?

7.13. Can the pair of amicable numbers 1184 and 1210 be constructed from Thabit ibn-Qurra's formula?

7.14. Solve the generalized problem stated by Matsunaga of finding an integer N that is simultaneously of the form $x^2 + a_1x + b_1$ and $y^2 + a_2y + b_2$. To do this, show that it is always possible to factor the number $(a_2^2 + 4b_1) - (a_1^2 + 4b_2)$ as a product mn , where m and n are either both even or both odd, and that the solution is found by taking $x = \frac{1}{2}(\frac{m-n}{2} - a_1)$, $y = \frac{1}{2}(\frac{m+n}{2} - a_2)$.

7.15. Leonardo's solution to the problem of finding a second pair of squares having a given sum is explained in general terms, then illustrated with a special case. He considers the case $4^2 + 5^2 = 41$. He first finds two numbers (3 and 4) for which the sum of the squares is a square. He then forms the product of 41 and the sum of the squares of the latter pair, obtaining $25 \cdot 41 = 1025$. Then he finds two squares whose sum equals this number: 31^2 and 8^2 or 32^2 and 1^2 . He thus obtains the results $(\frac{31}{5})^2 + (\frac{8}{5})^2 = 41$ and $(\frac{32}{5})^2 + (\frac{1}{5})^2 = 41$. Following this method, find another pair of rational numbers whose sum is 41. Why does the method work?

7.16. If the general term of the Fibonacci sequence is a_n , show that $a_n < a_{n+1} < 2a_n$, so that the ratio a_{n+1}/a_n always lies between 1 and 2. Assuming that this ratio has a limit, what is that limit?

7.17. Suppose that the pairs of rabbits begin to breed in the *first* month after they are born, but die after the second month (having produced two more pairs). What sequence of numbers results?

7.18. Prove that if x , y , and z are relatively prime integers such that $x^2 + y^2 = z^2$, with x and z odd and y even, there exist integers u and v such that $x = u^2 - v^2$,

$y = 2uv$, and $z = u^2 + v^2$. [*Hint*: Start from the fact that $x^2 = (z - y)(z + y)$, so that $z - y = a^2$ and $z + y = b^2$ for some a and b .]

CHAPTER 8

Numbers and Number Theory in Modern Mathematics

Beginning with the work of Fermat in the seventeenth century, number theory has become ever more esoteric and theoretical, developing connections with algebra and analysis that lie very deep and require many years of study to master. Obviously, we cannot explain in any satisfactory detail what has happened in this area in recent years. For that reason, we shall carry the story forward only as far as the beginning of the twentieth century.

A second topic that we must discuss before leaving the subject of numbers is the variety of invented number systems, starting with the natural positive integers. The number zero, negative numbers, and rational numbers do not require a long explanation, but we need to focus in more detail on real and complex numbers and the cardinal and ordinal arithmetic that came along with set theory.

Finally, mere counting turns out to be very difficult in some cases; for example, given twelve points on a circle, each pair of which is joined by a chord, into how many regions will these chords divide the circle if no three chords intersect in a common point? To solve such problems, sophisticated methods of counting have been developed, leading to the modern subject of combinatorics. A survey of its history concludes our study of numbers.

1. Modern number theory

We are forced to leave out many important results in our survey of modern number theory. Dickson's summary of the major results (1919, 1920, 1923) occupies 1600 pages, and an enormous amount of work has been added since it was published. Obviously, the present discussion is going to be confined to a few of the most significant authors and results.

1.1. Fermat. Pierre de Fermat (1603–1665) was a lawyer in Toulouse whose avid interest in mathematics led him to create, in his spare time, some analytic geometry, calculus, and modern number theory. According to one source book (Smith, 1929, p. 214), he was “the first to assert that the equation $x^2 - Dy^2 = 1$ has infinitely many solutions in integers.” As we have seen, given that it has *one* solution in integers, Brahmagupta knew 900 years before Fermat that it must have infinitely many, since he knew how to create new solutions from old ones. It was mentioned above that Fermat wrote in the margin of his copy of Diophantus that the sum of two positive rational cubes could not be a rational cube, and so on (Fermat's last theorem). Although Fermat never communicated his claimed proof of this fact, he was one of the first to make use of a method of proof—the method of infinite descent—by which many facts in number theory can be proved, including the case of fourth powers in Fermat's last theorem. A proof of the case of fourth powers

was given by Euler in 1738. Actually, the proof shows that there can be no positive integers x, y, z such that $x^4 + y^4 = z^2$.

Another area of number theory pioneered by Fermat arises naturally from consideration of quadratic Diophantine equations. The question is: "In how many ways can an integer be represented as a sum of two squares?" A number of the form $4n + 3$ cannot be the sum of two squares. This is an easy result, since when a square is divided by 4 the remainder is either 0 or 1, and it is impossible to write 3 as the sum of two numbers, each of which is either 0 or 1. But Fermat proved the much more difficult result that a prime number of the form $4n + 1$ can be written as the sum of two squares in exactly one way. Thus, $73 = 8^2 + 3^2$, for example.

The work of Fermat in number theory was continued by many mathematicians in the eighteenth century. We shall discuss very briefly the lives and work of three of them.

1.2. Euler. The Swiss mathematician Leonhard Euler (1707–1783) was one of the most profound and prolific mathematical writers who ever lived, despite having lost the sight of one eye early in life and the other later on. His complete works have only recently been assembled in good order. In 1983, the two hundredth anniversary of his death, many memorial volumes were dedicated to him, including an entire issue (45, No. 5) of *Mathematics Magazine*. An even larger celebration is planned for 2007, the three hundredth anniversary of his birth.¹ He spent most of the years from 1726 to 1741 and from 1766 until his death in St. Petersburg, Russia, where he was one of the first members of the Russian Academy of Sciences, founded by Tsar Peter I (1672–1725) just before his death. From 1741 to 1766 he was in Berlin, at the Prussian Academy of Sciences of Frederick II (1712–1788). The exact date at which he made many of his great discoveries is sometimes difficult to establish, and different dates are sometimes given in the literature. Euler's contributions to the development of calculus, differential equations, algebra, geometry, and mathematical physics are enormous. The following paragraphs describe some of his better-known results in number theory.

Fermat primes. Fermat had conjectured that the number $F_n = 2^{(2^n)} + 1$ is always a prime. This statement is true for $n = 0, 1, 2, 3, 4$, as the reader can easily check. For $n = 5$ this number is 4,294,967,297, and to prove that it is prime using the sieve of Eratosthenes, one must attempt to divide it by every prime less than $F_4 = 65,537$. In 1732 Euler found that this fifth Fermat number is divisible by 641. No Fermat number beyond F_4 has ever been shown to be prime, and well over 200 are now known to be composite, including $F_{2478782}$, discovered by John Cosgrave and others at St. Patrick's College, Dublin, on October 10, 2003. The smallest Fermat number not definitely known to be either prime or composite is F_{33} . The problem of Fermat primes is almost, but not quite, an idle question, that is, one without connections to anything else in mathematics. The connection in this case is that the regular polygons that have an odd number of sides and can be constructed with straightedge and compass are precisely those whose number of sides is a product of Fermat primes. Thus, until such time as another Fermat number is proved to be prime, the only Euclidean-constructible regular polygons

¹ See the website <http://www.euler2007.com/>.

with an odd number of sides will be those whose number of sides is a product of distinct numbers from the following short list: 3, 5, 17, 257, 65,537.²

Fermat's last theorem: the cases $n = 4$ and $n = 3$. We pointed out above that Fermat's method of infinite descent can be used to prove Fermat's last theorem for $n = 4$, as shown by Euler in 1738. In a textbook of algebra published in 1770 (see Struik, 1986, pp. 36–40), Euler also gave a proof of the impossibility for the case $n = 3$, which is much more difficult. In 1772 Euler proved that every positive integer is the sum of at most four square integers and conjectured that no sum of fewer than n n th powers could be an n th power, a conjecture that was finally refuted for $n = 5$ in 1966.³

Fermat's little theorem. A second assertion of Fermat, which Euler proved in 1736 (see Struik, 1986, p. 35), is known as *Fermat's little theorem*. It asserts that if p is a prime number that does not divide a number a , then p does divide $a^d - 1$ for some positive integer d ; moreover, the smallest d for which this statement is true divides $p - 1$. In particular, p divides $a^{p-1} - 1$. Fermat's discovery of this theorem has an interesting history (Fletcher, 1989). Through Mersenne Fermat had received a challenge in 1640 from Bernard Frénicle de Bessy (ca. 1612–1675) to find the first perfect number having at least 20 digits. It was known that $2^{30}(2^{31} - 1)$ was a perfect number having 19 digits.⁴ Fermat did not succeed in doing this; but after studying the problem, he noted that if n is composite, then $2^n - 1$ is also and that if n is a prime, then $2^n - 2$ is divisible by $2n$, that is, n divides $2^{n-1} - 1$. Moreover, he said, all other prime divisors of $2^n - 2$ leave a remainder of 1 when divided by $2n$. These theorems, in Fermat's view, were the secret of discovering perfect numbers, and he asserted that there were none having 20 or 21 digits. In a letter to Frénicle in October 1640, Fermat stated his "little" theorem—so called to distinguish it from his greater, "last" theorem. It is by no means a "little" result. Euler extended this result and showed that m divides $a(a^{\phi(m)} - 1)$, where $\phi(m)$ is the number of positive integers less than m and relatively prime to m (now called *Euler's ϕ -function*). That function provides the theoretical basis for constructing the RSA⁵ codes that are an essential part of communications security. Thus, the completely "useless" topic of perfect numbers actually inspired a number-theoretic result of great practical value.

Residues modulo a prime. Euler defined an integer n to be a λ -power residue with respect to ("modulo," as we now say, using Euler's Latin term) a prime p if there is an integer a such that p divides $a^\lambda - n$. This concept has proved to be a rich source of investigation in number theory. In particular, the case $\lambda = 2$ (quadratic residues) has led to some deep theorems. In works published in 1751 and 1783,

² Heinrich Wefelscheid informs me that one Johann Hermes (1846–1912), a student in Königsberg from 1866 to 1870, actually attempted to work out the case 65,537 and published his method in the *Göttinger Nachrichten* in 1894. The Australian mathematician Joan Taylor has used a computer to complete Hermes' project and found that the algebraic expression for $\cos(2\pi/65537)$ occupies 12.5 megabytes and contains an integer of 19,717 digits.

³ In connection with this result, we note that Fermat had stated a positive conjecture: *Every positive integer is the sum of at most n n -gonal numbers, that is, three triangular numbers, four squares, five pentagonal numbers, and so on.* This result was first proved for the general case by Augustin-Louis Cauchy (1789–1856) in 1813.

⁴ Specifically, it is 2,305,843,008,139,952,128.

⁵ Invented in 1977 and named from the initials of its three inventors, Ronald L. Rivest, Adi Shamir, and Leonard Adleman.

Euler conjectured what we now know as the law of *quadratic reciprocity*: *Given two primes p and q both of which equal 3 modulo 4, exactly one of them is a quadratic residue modulo the other. In all other cases, either each is a quadratic residue modulo the other or neither is.* For example, $11 \equiv 2^2 \equiv 5^2 \pmod{7}$, but the quadratic residues modulo 11 are 0, 1, 4, 3 ($\equiv 5^2 \equiv 6^2 \pmod{11}$), 5 ($\equiv 4^2 \equiv 7^2 \pmod{11}$), and 9; 7 is not among them. That is because both 7 and 11 are equal to 3 modulo 4. On the other hand, since 5 equals 1 modulo 4, we find that $11 \equiv 1^2 \pmod{5}$ and $5 \equiv 7^2 \pmod{11}$; similarly, neither 5 nor 7 is a quadratic residue modulo the other. The fact that Euler did not succeed in proving the law of quadratic reciprocity shows how difficult a result it is.

The Goldbach conjecture. A problem of number theory whose fame is second only to the Fermat conjecture is a conjecture of Christian Goldbach (1690–1764), who wrote to Euler in 1742 that every integer seemed to be a sum of at most three prime integers (Struik, 1986, pp. 47–49). Euler wrote back that he believed, but was unable to prove, the stronger proposition that every even integer larger than 4 is the sum of two odd primes—in other words, that one of the three primes conjectured by Goldbach can be chosen arbitrarily. Euler’s statement is known as the *Goldbach conjecture*. In 1937 the Russian mathematician Ivan Matveevich Vinogradov (1891–1983) proved that every sufficiently large odd integer is the sum of at most three primes.

1.3. Lagrange. The generation after Euler produced the Italian–French mathematician Joseph-Louis Lagrange (1736–1813). His name gives the impression that he was French, and indeed his ancestry was French and he wrote in French; but then so did many others, as French was literally the “lingua franca,” the common language of much scientific correspondence during the eighteenth and nineteenth centuries. Lagrange was born in Turin, however, and lived there for the first 30 years of his life, signing his first name as “Luigi” on his first mathematical paper in 1754. When the French Revolution came, he narrowly escaped arrest as a foreigner; and we have the word of Jean-Joseph Fourier, who heard him lecture, that he spoke French with a noticeable Italian accent. Thus it appears that the Italians are correct in claiming him as one of their own, even though his most prominent works were published in France and he was a member of the Paris Academy of Sciences for the latter part of his life.

Lagrange’s early work impressed Euler, then in Berlin, very favorably, and attempts were made to bring him to Berlin. But the introverted Lagrange seems to have been intimidated by Euler’s power as a mathematician and refused all such offers until Euler went back to St. Petersburg in 1766. He then came to Berlin and remained there until the death of Frederick II in 1788, at which point he accepted a position at the Paris Academy of Sciences, where he spent the last 15 years of his life. Lagrange did important work in algebra and mechanics that is discussed in later chapters. At this point we note only some of his number-theoretic results.

The Pell equation. Shortly after arriving in Berlin in 1766, Lagrange gave a definitive discussion of the solutions of the Pell equation $x^2 = Dy^2 \pm 1$, using the theory of continued fractions. In the course of this work he proved the important fact that any irrational number satisfying a quadratic equation with integer coefficients has a periodic continued fraction expansion. The converse of that statement is also true, and it turns out that the continued-fraction expansion of \sqrt{D} can be used to

construct *all* solutions of the Pell equation $Dx^2 \pm 1 = y^2$ (see Scharlau and Opolka, 1985, pp. 45–56).

The four-squares theorem. In 1770 Lagrange gave a proof that every integer is the sum of at most four square integers (which Euler also proved a year or so later).

“Wilson’s theorem”. In 1771 Lagrange proved that an integer n is prime if and only if n divides $(n-1)! + 1$. Thus 5 is prime because $4! + 1 = 25$, but 6 is not prime because it does not divide $5! + 1 = 121$. This theorem was attributed to John Wilson (1741–1793) by the Cambridge professor Edward Waring (1736–1798), who was apparently unaware that it was first stated by al-Haytham (965–1040). No proof of it can be found in the work of Wilson, who left mathematics to become a lawyer.

Quadratic binary forms. The study of quadratic Diophantine equations involves expressions of the form $ax^2 + bxy + cy^2$. The integers that can be represented in this way for given values of a , b , and c were the subject of two memoirs by Lagrange, amounting to nearly 100 pages of work, during the years 1775–1777.

1.4. Legendre. The volume of work on number theory increased greatly in the last half of the eighteenth century, and the first treatises devoted specifically to that subject appeared. One of the prominent figures in this development was Adrien-Marie Legendre (1752–1833). Like all other mathematicians of the time, Legendre worked in many areas of mathematics, including calculus (elliptic functions) and mechanics. He also worked in number theory and produced several profound results there in an early textbook of the subject, which went through three editions before his death.

In 1785 he published the paper “Recherches d’analyse indéterminée,” in which he proved the elegant result that there are integers x, y, z satisfying an equation $ax^2 + by^2 + cz^2 = 0$ with a, b, c not all of the same sign if and only if the products $-ab$, $-bc$, and $-ca$ are quadratic residues modulo $|c|$, $|a|$, and $|b|$ respectively. He also stated the law of quadratic reciprocity, which Euler had been unable to prove, and gave a flawed proof of it. He invented the still-used Legendre symbol $\left(\frac{p}{q}\right)$ whose value is 1 if p is a quadratic residue modulo q and -1 if not. The law of quadratic reciprocity can then be elegantly stated as $\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{(p-1)(q-1)}{4}}$. This proof was improved in his treatise *Théorie des nombres*, published in 1798, with a subsequent edition in 1808 and a third in 1830. He also conjectured, but did not prove, that any arithmetic sequence in which the constant difference is relatively prime to the first term will contain infinitely many primes. In fact, it was this unproved assumption that invalidated his proof of quadratic reciprocity (see Weil, 1984, pp. 329–330). He quoted Fermat’s conjecture that every number is the sum of at most n n -gonal numbers, noting with regret that either Fermat never completed the treatise he intended to write or that his executors never found the manuscript. Legendre gave a proof of this fact for all numbers larger than $50n - 79$. Further continuing the work of Fermat, Euler, and Lagrange, Legendre discovered some important facts in the theory of quadratic forms.

His most original contribution to number theory, however, lay in a different direction entirely. Since no general law had been found for describing the n th prime number or even producing a polynomial whose values are all prime numbers, Legendre’s attempt to estimate the number of primes among the first n integers, published in the second (1808) edition of *Théorie des nombres*, was an important

step forward. Legendre's estimate for this number, which is now denoted $\pi(n)$ (π for *prime*, of course), was

$$\pi(n) \approx \frac{n}{\log n - 1.08366} = \frac{n}{\log n} \left(1 + \frac{1.08366}{\log n} + \frac{1.08366^2}{\log^2 n} + \cdots \right).$$

Here the logarithm is understood as the natural logarithm, what calculus books usually denote $\ln n$. In particular, the ratio $\frac{\pi(n)}{n/\log n}$ tends to 1 as n tends to infinity. Legendre did not have a proof of this result, but merely conjecturing it was an important advance in the understanding of primes.

Legendre also worked on the classification of real numbers; his contributions to this area are described below.

Number theory blossomed in the nineteenth century due to the attention of many brilliant mathematicians. Again, we have space to discuss only a few of the major figures.

1.5. Gauss. Carl Friedrich Wilhelm Gauss (1777–1855), one of the giants of modern mathematics, lived his entire life in Germany. He studied at the University of Göttingen from 1795 to 1798 and received a doctoral degree in 1799 from the University of Helmstedt. Thereafter most of his life was spent in and around Göttingen, where he did profound work in several areas of both pure and applied mathematics. In particular, he worked in astronomy, geodesy, and electromagnetic theory, producing fundamental results on the use of observational data (least squares), mapping (Gaussian curvature), and applied electromagnetism (the telegraph). But his results in pure number theory are among the deepest ever produced. Here we look at just a few of them.



Street in Göttingen named after Gauss.

The 1801 work *Disquisitiones arithmeticae* became a classical work on the properties of integers. One the earliest discoveries that Gauss made, when he was still a teenager, was a proof of the law of quadratic reciprocity. This proof was published in the *Disquisitiones*, and over the next two decades he found seven more proofs of this fundamental fact. The *Disquisitiones* also contain a proof of the fundamental theorem of arithmetic and a construction of the regular 17-sided polygon, which is possible because 17 is a Fermat prime.

A considerable portion of the *Disquisitiones* is devoted to quadratic binary forms, in an elegant and sophisticated treatment that contemporaries found difficult to understand. As Weil says (1984, p. 354),

No doubt the Gaussian theory... is far more elaborate [than Legendre's treatment of the subject]; so much so, indeed, that it remained a stumbling-block for all readers of the *Disquisitiones*

until Dirichlet restored its simplicity by going back very nearly to Legendre's original construction.

In attempting to extend the law of quadratic reciprocity to higher powers, Gauss was led to consider what are now called the *Gaussian integers*, that is, complex numbers of the form $m + n\sqrt{-1}$. Gauss showed that the concepts of prime and composite number make sense in this context just as in the ordinary integers and that every such number has a unique representation (up to multiplication by the units ± 1 and $\pm\sqrt{-1}$) as a product of irreducible factors. Notice that no prime integer of the form $4n + 1$ can be "prime" in this context, since it is a sum of two squares: $4n + 1 = p^2 + q^2 = (p + q\sqrt{-1})(p - q\sqrt{-1})$. The generalization of the notion of prime number to the Gaussian integers is an early example of the endless generalization and abstraction that characterizes modern mathematics.

Gauss also gave an estimate of the number of primes not larger than x , in the form of the integral

$$\pi(x) \approx \text{Li}(x) = \int_2^x \frac{dt}{\log(t)}.$$

Here, as above, the logarithm means the natural logarithm. He did not, however, prove that this approximation is asymptotically good, that is, that $\pi(x)/\text{Li}(x)$ tends to 1 as x tends to infinity. That is the content of the prime number theorem.

1.6. Dirichlet. The works of Gauss on number theory were read by another bright star of nineteenth-century mathematics, Johann Peter Gustav Lejeune-Dirichlet (1805–1859), who contributed several gems to this difficult area. He was of Belgian ancestry (hence his French-sounding name, even though he was a German). He was born in the city of Düren, which lies between Aachen (Aix) and Köln (Cologne), but went to Paris to study at the age of 16. At the age of 20 he proved the case $n = 5$ of Fermat's last theorem. (Legendre, who was the referee for Dirichlet's paper, contributed his own proof of one subcase of this case.) That same year he returned to Germany and took up a position at the University of Breslau. In 1828 he went to Berlin and was the first star in a bright galaxy of Berlin mathematicians. In 1831 he was elected to the Berlin Academy of Sciences. That year he married Rebekah Mendelssohn, sister of the composers Felix and Fanny Mendelssohn. In 1855, dissatisfied with the heavy teaching loads in Berlin, he moved to Göttingen as the successor of Gauss, who had died that year. In 1858 he suffered a heart attack and the death of his wife, and in 1859 he himself succumbed to heart disease.

Although Dirichlet also worked in the theory of Fourier series and analytic function theory, having given the first rigorous discussion of the convergence of a Fourier series in 1829 and the modern definition of a function in 1837, we are at the moment concerned with his contributions to number theory. One of these is his 1837 theorem, already mentioned, that each arithmetic sequence in which the first term and the common difference are relatively prime contains an infinite number of primes. To prove this result, he introduced what is now called the *Dirichlet character* $\chi(n) = (-1)^k$ if $n = 2k + 1$, $\chi(n) = 0$ if n is even, along with the *Dirichlet series*

$$\sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} = 1 - \frac{1}{3^s} + \frac{1}{5^s} - \frac{1}{7^s} + \cdots.$$

This work brought number theory and analysis together in the subject now called analytic number theory. According to Weil (1984, pp. 252–256), the two

subjects had been drawing closer together ever since Euler began his study of elliptic functions. Elliptic functions have played a prominent role in number theory since 1830. Carl Gustav Jacob Jacobi (1804–1851), whose work is discussed in Chapter 17, published a treatise on elliptic functions in 1829 in which he used these functions to derive a formula equivalent to

$$(1) \quad \left(\sum_{n=-\infty}^{+\infty} q^{n^2} \right)^4 = 1 + 8 \sum_{k=1}^{\infty} \sigma(2k-1) q^{2k-1} + 24 \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \sigma(2k-1) q^{2^j(2k-1)},$$

where $\sigma(r)$ is the sum of the divisors of r . For $r > 0$, it is obvious that the coefficient of q^r on the left is 16 times the number of ways in which r can be represented as the sum of four ordered nonzero squares, plus 32 times the number of such representations as a sum of three squares, plus 24 times the number of representations as a sum of two squares, plus 8 if the number happens to be a square. The coefficient of q^r on the right is either eight or 24 times the sum of the divisors of the largest odd number that divides r . Since that sum is always positive, the four-square theorem is a consequence, but much additional information is added on the number of such representations.

The study of Dirichlet series, in particular the simplest one of all, which defines what is now called the *Riemann zeta function*

$$\zeta(z) = \sum_{n=1}^{\infty} \frac{1}{n^z}$$

(one of several zeta functions named after distinguished mathematicians), turned out to be important in both complex analysis and number theory. The zeta function was introduced, though not under that name, by Euler, who gave the formula

$$(2) \quad \sum_{n=1}^{\infty} \frac{1}{n^z} = \prod_{p \text{ prime}} \left(1 - \frac{1}{p^z} \right).$$

The fact that the terms in the sum are indexed by all positive integers while the factors in the product are indexed by the prime numbers accounts for the deep connections of this function with number theory. Its values at the even integers can be computed in terms of the Bernoulli numbers.⁶ In fact, the Bernoulli numbers were originally introduced this way. Nowadays, the n th Bernoulli number B_n is defined to be $n!$ times the coefficient of x^n in the Maclaurin series of $x/(e^x - 1)$.

1.7. Riemann. Another giant of nineteenth-century mathematics was Georg Bernhard Riemann (1826–1866), who despite his brief life managed to make major contributions to real and complex analysis, geometry, algebraic topology, and mathematical physics. He was inspired by Legendre's work on number theory and studied under Gauss at Göttingen, where he also became a professor, Dirichlet's successor after 1859. Because of his frail health (he succumbed to tuberculosis at the age of 40), he spent considerable time in Italy, where he made the acquaintance of the productive school of Italian geometers, including Enrico Betti (1823–1892). His greatest contribution to number theory was to attempt a rigorous estimate of $\pi(n)$. For this purpose he studied the zeta function introduced above and made the famous conjecture that except for its obvious zeros at the even negative integers, all

⁶ The Bernoulli numbers were the object of the first computer program written for the Babbage analytical engine.

other zeros have real part equal to $\frac{1}{2}$. This *Riemann hypothesis* forms one of the still-outstanding unsolved problems of modern mathematics, standing alongside the Goldbach conjecture and a famous conjecture in topology due to Henri Poincaré (1854–1912).⁷ The first two were mentioned by Hilbert in his address at the 1900 International Congress of Mathematicians in Paris. Hilbert gave the Riemann hypothesis as the eighth of his list of 23 unsolved problems and suggested that solving it would also solve the Goldbach conjecture. Despite a great many partial results, the complete problem remains open a century later. A summary of the work on this problem through the mid-twentieth century can be found in the book by Edwards (1974). The zeros of $\zeta(z)$ are now being computed at a furious rate by the ZetaGrid project, an Internet-based distributed program linking tens of thousands of computers, similar to the GIMPS mentioned above (see www.zetagrid.net).



Poster of Riemann at the Mathematisches Institut and street named after him in Göttingen.

1.8. Fermat's last theorem. Work on Fermat's most famous conjecture continued in the nineteenth century. In 1847 Gabriel Lamé (1795–1870), published a paper in which he claimed to have proved the result. Unfortunately, he assumed

⁷ This conjecture may now have been solved (see Section 4 of Chapter 12).

that complex numbers of the form $a_0 + a_1\theta + \cdots + a_{n-1}\theta^{n-1}$, where $\theta^n = 1$ and a_0, \dots, a_{n-1} are integers, can be factored uniquely, just like ordinary integers. Ernst Eduard Kummer (1810–1893) had noticed some 10 years earlier that such is not the case. This was just one of the many ways in which the objects studied by mathematicians became increasingly abstract, and the old objects of numbers and space became merely special cases of the general objects about which theorems are proved. Kummer was the first to make general progress toward a proof of Fermat's last theorem. The conjecture that $x^p + y^p = z^p$ has no solutions in positive integers x, y , and z when p is an odd prime had been proved only for the cases $p = 3, 5$, and 7 until Kummer showed that it was true for a class of primes called *regular primes*, which included all the primes less than 100 except 37, 59, and 67. This step effectively closed off the possibility that Fermat might be proved wrong by calculating a counterexample.

1.9. The prime number theorem. A good estimate of the number of primes less than or equal to a given integer N is given by $N/(\log N)$. This estimate follows from the unproved estimate of Gauss given above. The estimate suggested by Legendre, $N/(A \log N + B)$ with $A = 1$, $B = -1.08366$, turns out to be correct only in its first term. This fact was realized by Dirichlet, but only after he had written approvingly of the estimate in print. (He corrected himself in a marginal note on a copy of his paper given to Gauss.) Dirichlet suggested $\sum_{k=2}^N (1/(\log k))$ as a better approximation. This problem was also studied by the Russian mathematician Pafnutii L'vovich Chebyshev (1821–1894).⁸ In 1851 Chebyshev proved that if $\alpha > 0$ is any positive number (no matter how small) and m is any positive number (no matter how large), the inequality

$$\pi(n) > \int_2^n \frac{dx}{\log x} - \frac{\alpha n}{\log^m n}$$

holds for infinitely many positive integers n , as does the inequality

$$\pi(n) < \int_2^n \frac{dx}{\log x} + \frac{\alpha n}{\log^m n}.$$

This result suggests that $\pi(n) \sim [n/(\ln n)]$, but it would be desirable to know if there is a constant A such that

$$\pi(n) = \frac{An}{\log n} + \varepsilon_n,$$

where ε_n is of smaller order than $\pi(n)$. It would also be good to know the rate at which $\varepsilon_n/\pi(n)$ tends to zero. Chebyshev's estimates imply that if A exists, it must be equal to 1, and as a result, Legendre's approximation cannot be valid beyond the first term. Chebyshev was able to show that

$$0.92129 < \frac{\pi(n)}{\log n} < 1.10555.$$

Chebyshev mentions only Legendre in his memoir on this subject and shows that his estimates refute Legendre's conjecture. He makes no mention of Gauss, whose integral $\text{Li}(x)$ appears in his argument. Similarly, Riemann makes no mention of

⁸ In Russian this name is pronounced "Cheb-wee-SHAWF," approximately. However, because he wrote so often in French, where he signed his name as "Tchebycheff," it is usually pronounced "CHEB-ee-shev" in the West.

Chebyshev in his 1859 paper on $\pi(x)$, even though he was in close contact with Dirichlet, and Chebyshev's paper had been published in a French journal.

The full proof of the prime number theorem turned out to involve the use of complex analysis. As mentioned above, Riemann had studied the zeros of the Riemann zeta function. This function was also studied by two long-lived twentieth-century mathematicians, the Belgian Charles de la Vallée Poussin (1866–1962) and the Frenchman Jacques Hadamard (1865–1963), who showed independently of each other (Hadamard, 1896; Vallée Poussin, 1896) that the Riemann zeta function has no zeros with real part equal to 1.⁹ Vallée Poussin showed later (1899) that

$$\pi(n) = \int_2^n \frac{dx}{\log x} + \varepsilon_n,$$

where for some $\alpha > 0$ the error term ε_n is bounded by a multiple of $ne^{-\alpha\sqrt{\log n}}$.

Number theory did not slow down or stop after the proof of the prime number theorem. On the contrary, it exploded into a huge number of subfields, each producing a prodigious amount of new knowledge year by year. However, we must stop writing on this subject sometime and move on to other topics, and so we shall close our account of number theory at this point.

2. Number systems

To the ancient mathematicians in the Middle East and Europe, numbers meant positive integers or ratios of them, in other words, what we call rational numbers. In India and China negative numbers were recognized, and 0 was recognized as a number in its own right, as opposed to merely an absence of numbers, at a very early stage. Those numbers reached Europe only a brief while before algebra led to the consideration of imaginary numbers. In this section we explore the gradual expansion of the concept of a number to include not only negative and imaginary numbers, which at least had the merit of being understandable in finite terms, but also irrational roots of equations and transcendental real numbers such as π and e , and the infinite cardinal and ordinal numbers mathematicians routinely speak about today. It is a story of the gradual enlargement of the human imagination and the clarification of vague, intuitive ideas.

2.1. Negative numbers and zero. It was mentioned in Chapter 7 that place-value systems of writing numbers were invented in Mesopotamia, India, China, and Mesoamerica. What is known about the Maya system has already been described in Chapter 5. We do not know how or even if they performed multiplication or division or how they worked with fractions. Thus, for this case all we know is that they had a place-value system and that it included a zero to occupy empty places. The Mesopotamian system was sexagesimal and had no zero for at least the first 1000 years of its existence. The other three systems were decimal, and they too were rather late in acquiring the zero. Strange though it may seem to one who has a modern education, in India and China negative numbers seem to have been used before zero was invented.

⁹ The fact that $\zeta(z)$ has no zeros with real part equal to 1 is an elementary theorem (see Ivić, 1985, pp. 7–8). That does not make the prime number theorem trivial, however, since the equivalence between this result and the prime number theorem is very difficult to prove. A discussion of the reasons why the two are equivalent was given by Norbert Wiener; see his paper "Some prime-number consequences of the Ikehara theorem," in his *Collected Works*, Vol. 2, pp. 254–257.

China. Chinese counting rods were red or black according as the numbers represented were positive or negative, yet no zero occurs, since obviously it would be absurd to have a rod representing no rods at all. A Chinese work on astronomy and the calendar written in the late second century CE (Li and Du, 1987, p. 49) gives rules for adding and subtracting “strong” (positive) and “weak” (negative): When adding, like signs add and opposite signs subtract; when subtracting, like signs subtract and opposite signs add. The same kinds of rules are given in Chapter 8 (“Rectangular Tabulation”) of the *Jiu Zhang Suanshu*. Yet it was a full thousand years after that time when the rules for multiplying and dividing signed numbers first appeared in the *Suanshu Chimeng* of 1303. When one is using a counting board or an abacus, no symbol for zero is needed, since it is visually apparent that a given square has no numbers written in it or that the beads on a string are in their “empty” position. The first known occurrence of the symbol 0 for zero occurs in a work dating to the year 1247.

India. Around the year 500, Aryabhata I, used a place-value decimal system without zero. A century later Brahmagupta introduced zero in connection with the *kuttaka* method described above. Although he used the word *sunya* (empty) for this concept and it really does denote an empty place in that method, the idea that the algorithm $x \mapsto ax + b$ can be executed as $x \mapsto ax$ when no b is present suggests the use of a neutral element for addition, and that is what the zero is. Brahmagupta gave complete rules for addition, subtraction, multiplication, and division of both positive and negative quantities and zero. As we know, division by zero must be considered separately and either rejected or given some special meaning. Brahmagupta (Colebrooke, 1817, pp. 339–340) showed some puzzlement about this, and he wrote:

Cipher, divided by cipher, is nought. Positive, divided by negative, is negative. Negative, divided by affirmative, is negative. Positive or negative, divided by cipher, is a fraction with [cipher] for denominator, or cipher divided by negative or affirmative [is a fraction with the latter for denominator].

The word *cipher* here translates the Sanskrit *sunya* or *kha*, both meaning empty space. The last rule given here is not a happy effort at a definition; it is rather like saying that a jar contains its contents. Not much new information is conveyed by the sentence. But the obscurity is natural due to the complete absence of any human experience with situations corresponding to division by zero. Five hundred years later Bhaskara was still having trouble with this concept (Colebrooke, 1817, p. 19):

A definite quantity divided by cipher is the submultiple of nought [that is, a fraction with zero for its denominator, just as Brahmagupta had said]. The product of cipher is nought: but it must be retained as a multiple of cipher, if any further operation impend. Cipher having become a multiplier, should nought afterwards become a divisor, the definite quantity must be understood to be unchanged.

Although these principles might be more clearly stated, it seems that Bhaskara may have in mind here some operations similar to those that occur in limiting

operations, for example, considering the appropriate value of a fraction such as $(5x^2 + 4x)/(3x^2 - 2x)$ when x becomes zero. One can formally cancel the x without thinking about whether or not it is zero. After cancellation, setting $x = 0$ gives the fraction the value -2 . Bhaskara is explicit in saying that zero added to any number leaves that number unchanged. Hence for him it is more than a mere placeholder; arithmetic operations can be performed with it. The use of an empty circle or a circle with its center marked as a symbol for zero seems to be culturally invariant, since it appears in inscriptions in India from the ninth century, in Greek documents from the second century, and in Chinese documents from the thirteenth century.

Islamic number systems. The transmission of Hindu treatises to Baghdad led ultimately to the triumph of the numerals used today. According to al-Daffa (1973, p. 51) the Sanskrit words for an empty place were translated as the Arabic word *sifr*, which became the English words *cipher* and *zero* and their cognates in other European languages. Al-Daffa also points out that the earliest record of the symbol for zero in India comes from an inscription at Gwalior dating to 876, and that there is a document in Arabic dating from 873 in which this symbol occurs.

2.2. Irrational and imaginary numbers. In a peculiar way, the absence of a place-value system of writing numbers may have stimulated the creation of mathematics in ancient Greece in the case of irrational numbers. Place-value notation provides approximate square roots in practical form, even when the expansion does not terminate.¹⁰ A cuneiform tablet from Iraq (Yale Babylonian Collection 7289) shows a square with its diagonals drawn and the sexagesimal number 1;24,51,10, which gives the length of the diagonal of a square of side 1 to great precision. But in all the Chinese, Mesopotamian, Egyptian,¹¹ and Hindu texts there is nothing that can be considered a theoretical discussion of “numbers” whose expansions do not terminate.

The word *numbers* is placed in inverted commas here because the meaning of the square root of 2 is not easy to define. It is very easy to go around in circles making the definition. The difficulty came in a clash of geometry and arithmetic, the two fundamental modes of mathematical thinking. From the arithmetical point of view the problem is minimal. If numbers must be what we now call positive rational numbers, then some of them are squares and some are not, just as some integers are triangular, square, pentagonal, and so forth, while others are not. No one would be disturbed by this fact; and since the Greeks had no place-value system to suggest an infinite process leading to an exact square root, they might not have speculated deeply on the implications of this arithmetical distinction in geometry. But in fact, they did speculate on both the numerical and geometric aspects of the problem, as we shall now see. We begin with the arithmetical problem.

The arithmetical origin of irrationals: nonsquare rational numbers. In Plato's dialogue *Theatetus*, the title character reports that a certain Theodorus proved that the integers 2, 3, 5, and so on, up to 17 have no (rational) square roots, except of course the obvious integers 1, 4, and 9; and he says that for some reason, Theodorus stopped at that point. On that basis the students decided to classify numbers as

¹⁰ In the case of Chinese mathematics the end of a nonterminating square root was given as a common fraction, and Simon Stevin likewise terminated infinite decimals with common fractions.

¹¹ Square roots, called *corners*, are rarely encountered in the Egyptian papyri, and Gillings (1972, p. 214) suggests that they were found from tables of squares.

equilateral and *oblong*. The former class consists of the squares of rational numbers, for example $\frac{25}{9}$, and the latter are all other positive rational numbers, such as $\frac{3}{2}$.

One cannot help wondering why Theodorus stopped at 17 after proving that the numbers 3, 5, 6, 7, 8, 10, 11, 12, 13, 14, and 15 have no square roots. The implication is that Theodorus “got stuck” trying to prove this fact for a square of area 17. What might have caused him to get stuck? Most assuredly the square root of 17 is irrational, and the proof commonly given nowadays to show the irrationality of $\sqrt{3}$, for example, based on the unique prime factorization of integers, works just as well for 17 as for any other number. If Theodorus had our proof, he wouldn’t have been stuck doing 17, and he wouldn’t have bothered to do so many special cases, since the proofs are all the same. Therefore, we must assume that he was using some other method.

An ingenious conjecture as to Theodorus’ method was provided by Knorr (1945–1997) (1975). Knorr suggests that the proof was based on the elementary distinction between even and odd. To see how such a proof works, suppose that 7 is an equilateral number in the sense mentioned by Theatetus. Then there must exist two integers such that the square of the first is seven times the square of the second. We can assume that both integers are odd, since if both are even, we can divide them both by 2, and it is impossible for one of them to be odd and the other even. For the fact that the square of one of them equals seven times the square of the other would imply that an odd integer equals an even integer if this were the case. But it is well known that the square of an odd integer is always 1 larger than a multiple of 8. The supposition that the one square is seven times the other then implies that an integer 1 larger than a multiple of 8 equals an integer 7 larger than a multiple of 8, which is clearly impossible.

This same argument shows that none of the odd numbers 3, 5, 7, 11, 13, and 15 can be the square of a rational number. With a slight modification it can also be made to show that none of the numbers 2, 6, 8, 10, 12, and 14 is the square of a rational number, although no argument is needed in the case of 8 and 12, since it is already known that $\sqrt{2}$ and $\sqrt{3}$ are irrational. Notice that the argument fails, as it must, for 9: A number 9 larger than a multiple of 8 is also 1 larger than a multiple of 8. However, it also breaks down for 17 and for the same reason: A number 17 larger than a multiple of 8 is also 1 larger than a multiple of 8. Thus, even though it is *true* that 17 is not the square of a rational number, the argument just given, based on what we would call arithmetic modulo 8, cannot be used to *prove* this fact. In this way the conjectured method of proof would explain why Theodorus got stuck at 17.

The Greeks thus found not only that there was no integer whose square is, say, 11 (which is a simple matter of ruling out the few possible candidates), but also that there was not even any rational number having this property; that is, 11 is not the square of anything they recognized as a number.

The geometric origin of irrationals: incommensurable magnitudes. A second, “geometric” theory of the origin of irrational numbers comes from geometry and seems less plausible. If we apply the Euclidean algorithm to the side and diagonal of the regular pentagon in Fig. 1, we find that the pair AD and CD get replaced by lines equal to CD and CF , which are the diagonal and side of a smaller pentagon. Thus, no matter how many times we apply the procedure of the Euclidean algorithm, the result will always be a pair consisting of the side and diagonal of a

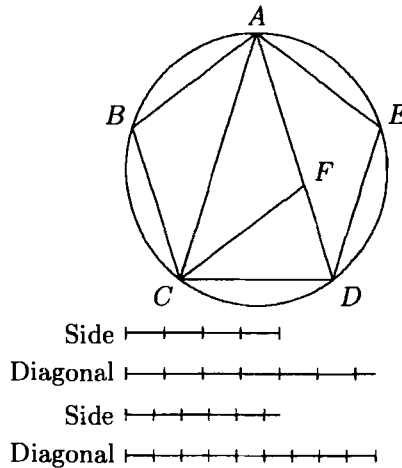


FIGURE 1. Diagonal and side of a regular pentagon. If a unit is chosen that divides the side into equal parts, it cannot divide the diagonal into equal parts, and vice versa.

pentagon. Therefore, in this case the Euclidean algorithm will *never* produce an equal pair of lines. We know, however, that it *must* produce an equal pair if a common measure exists. We conclude that *no common measure can exist for the side and diagonal of a pentagon*.

The argument just presented was originally given by von Fritz (1945). Knorr (1975, pp. 22–36) argues against this approach, however, pointing out that the simple arithmetic relation $d^2 = 2s^2$ satisfied by the diagonal and side of a square can be used in several ways to show that d and s could not both be integers, no matter what length is chosen as unit. Knorr prefers a reconstruction closer to the argument given in Plato's *Meno*, in which the problem of doubling a square is discussed.¹² Knorr points out that when discussing irrationals, Plato and Aristotle always invoke the side and diagonal of a square, never the pentagon or the related problem of dividing a line in mean and extreme ratio, which they certainly knew about.

Whatever the argument used, the Greeks discovered the existence of incommensurable pairs of line segments before the time of Plato. For Pythagorean metaphysics this discovery was disturbing: Number, it seems, is *not* adequate to explain all of nature. A legend arose that the Pythagoreans attempted to keep secret the discovery of this paradox.¹³ However, scholars believe that the discovery of incommensurables came near the end of the fifth century BCE, when the original Pythagorean group was already defunct.

¹² In Chapter 9 we invoke the same passage to speculate on the origin of the Pythagorean theorem.

¹³ The legend probably arose from a passage in Chapter 18, Section 88 of the *Life of Pythagoras* by Iamblichus. Iamblichus says that a certain Hippasus perished at sea, a punishment for his impiety because he published “the sphere of the 12 pentagons” (probably the radius of the sphere circumscribed about a dodecahedron), taking credit as if he had discovered it, when actually everything was a discovery of That Man (Pythagoras, who was too august a personage to be called by name). Apparently, new knowledge was to be kept in-house as a secret of the initiated and attributed in a mystical sense to Pythagoras.

The existence of incommensurables throws doubt on certain oversimplified proofs of geometric proportion, and this question is discussed in detail in Chapter 10. At present we are concerned with its effect on the concept of a *number*. At the beginning, one would have to say that the effect was almost nil. Geometry and arithmetic were separate subjects in the Greek tradition. But when algebra arose and the Persian mathematician Omar Khayyam (1050–1130) discovered that some cubic equations that could not be solved arithmetically had geometric solutions, the idea of a *real number* as a *ratio of lines* began to take shape.

The idea of using a line to stand for a number, the numbers being regarded as the length of the line, is very familiar to us and has its origin in the work of the ancient Greeks and medieval Muslim mathematicians. In Europe this idea received some development in the work of the fourteenth-century Bishop of Lisieux, Nicole d'Oresme, whose graphical representation of relationships was a forerunner of our modern analytic geometry. Oresme was familiar with the concept of incommensurable lines, a subject that was missing from earlier medieval work in geometry, and he was careful to keep the distinction between commensurable and incommensurable clear. Indeed, Oresme was even more advanced than the average twentieth-century person, in that he recognized a logical difficulty in talking about a power of, say, $\frac{1}{2}$ that equals $\frac{1}{3}$, whereas modern students are taught how to use the *rules* of exponents, but not encouraged to ask what is meant by expressions such as $\sqrt{2}^{\sqrt{3}}$.

A great advance came in the seventeenth century, when analytic geometry as we know it today was invented by Descartes and Fermat. Fermat's work seems somewhat closer to what we know, in the sense that he used a pair of mutually perpendicular axes; on the other hand, he believed that only dimensionally equivalent expressions could be added. This is the restriction that led Omar Khayyam to write a cubic equation in the form equivalent to $x^2 + ax^2 + b^2x = b^2c$, in which each term is of degree 3. In his *Géométrie*, Descartes showed how to avoid this complication. The difficulty lay in the geometric representation of the operation of multiplication. Because ratios of lines were not always numbers, Euclid did not make the association of a line with a number called its length. The product of two numbers is a number, but Euclid did not speak of the product of two lines. He spoke instead of the rectangle on the two lines. That was the tradition Omar Khayyam was following. Stimulated by algebra, however, and the application of geometry to it, Descartes looked at the product of two lengths in a different way. As pure numbers, the product ab is simply the fourth proportional to $1 : a : b$. That is, $ab : b :: a : 1$. He therefore fixed an arbitrary line that he called I to represent the number 1 and represented ab as the line that satisfied the proportion $ab : b :: a : I$, when a and b were lines representing two given numbers.

The notion of a *real number* had at last arisen, not as most people think of it today—an infinite decimal expansion—but as a ratio of line segments. Only a few decades later Newton defined a (real) number to be “the ratio of one magnitude to another magnitude of the same kind, arbitrarily taken as a unit.” Newton classified numbers as integers, fractions, and surds (Whiteside, 1967, Vol. 2, p. 7). Even with this amount of clarity introduced, however, mathematicians were inclined to gloss over certain difficulties. For example, there is an arithmetic rule according to which $\sqrt{ab} = \sqrt{a}\sqrt{b}$. Even with Descartes' geometric interpretation of these results, it is not obvious how this rule is to be proved. The use of the

decimal system, with its easy approximations to irrational numbers, soothed the consciences of mathematicians and gave them the confidence to proceed with their development of the calculus. No one even seemed very concerned about the absence of any good geometric construction of cube roots and higher roots of real numbers. The real line answered the needs of algebra in that it gave a representation of any real root there might be of any algebraic equation with real numbers as coefficients. It was some time before anyone realized that geometry still had resources that even algebra did not encompass and would lead to numbers for which pure algebra had no use.

Those resources included the continuity of the geometric line, which turned out to be exactly what was needed for the limiting processes of calculus. It was this property that made it sensible for Euler to talk about the number that we now call e , that is,

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = \sum_{n=0}^{\infty} \frac{1}{n!} = 2.7182818284590 \dots,$$

and the other Euler constant

$$\gamma = \lim_{n \rightarrow \infty} \left[\left(\sum_{k=1}^n \frac{1}{k} \right) - \log n \right] = 0.5772156649 \dots$$

The intuitive notion of continuity assured mathematicians that there were points on the line, and hence infinite decimal expansions, that must represent these numbers, even though no one would ever know the full expansions. The geometry of the line provided a geometric representation of real numbers and made it possible to reason about them without having to worry about the decimal expansion.

The continuity of the line brought the realization that the real numbers had more to offer than merely convenient representations of the solutions of equations. They could even represent some numbers such as e and γ that had not been found to be solutions of any equations. The line was richer than it needed to be for algebra alone. The concept of a real number had allowed arithmetic to penetrate into parts of geometry where even algebra could not go. The sides and diagonals of regular figures such as squares, cubes, pentagons, pyramids, and the like all had ratios that could be represented as the solutions of equations, and hence are algebraic. For example, the diagonal D and side S of a pentagon satisfy $D^2 = S(D + S)$. For a square the relationship is $D^2 = 2S^2$, and for a cube it is $D^2 = 3S^2$. But what about the number we now call π , the ratio of the circumference C of a circle to its diameter D ? In the seventeenth century Leibniz noted that any line that could be constructed using Euclidean methods (straightedge and compass) would have a length that satisfied some equation with rational coefficients. In a number of letters and papers written during the 1670s, Leibniz was the first to contrast what is algebraic (involving polynomials with rational coefficients) with objects that he called *analytic* or *transcendental* and the first to suggest that π might be transcendental. In the preface to his pamphlet *De quadratura arithmetica circuli* (*On the Arithmetical Quadrature of the Circle*), he wrote:

A complete quadrature would be one that is both analytic and linear; that is, it would be constructed by the use of curves whose equations are of [finite] degrees. The brilliant Gregory [James Gregory, 1638–1675], in his book *On the Exact Quadrature of the Circle*, has claimed that this is impossible, but, unless I am mistaken,

has given no proof. I still do not see what prevents the circumference itself, or some particular part of it, from being measured [that is, being commensurable with the radius], a part whose arc has a ratio to its sine [half-chord] that can be expressed by an equation of finite degree. But to express the ratio of the arc to the sine *in general* by an equation of finite degree is impossible, as I shall prove in this little work. [Gerhardt, 1971, Vol. 5, p. 97]

No representation of π as the root of a polynomial with rational coefficients was ever found. This ratio had a long history of numerical approximations from all over the world, but no one ever found any nonidentical equation satisfied by C and D . The fact that π is transcendental was first proved in 1881 by Ferdinand Lindemann (1852–1939). The complete set of real numbers thus consists of the positive and negative rational numbers, all real roots of equations with integer coefficients (the *algebraic* numbers), and the transcendental numbers. All transcendental numbers and some algebraic numbers are irrational. Examples of transcendental numbers are rather difficult to produce. The first number to be proved transcendental was the base of natural logarithms e , and this proof was achieved only in 1873, by the French mathematician Charles Hermite (1822–1901). It is still not known whether the Euler constant $\gamma \approx 0.57712$ is even irrational.

The arithmetization of the real numbers. Not until the nineteenth century, when mathematicians took a retrospective look at the magnificent edifice of calculus that they had created and tried to give it the same degree of logical rigor possessed by algebra and Euclidean geometry, were attempts made to define real numbers arithmetically, without mentioning ratios of lines. One such definition by Richard Dedekind (1831–1916), a professor at the Zürich Polytechnikum, was inspired by a desire for rigor when he began lecturing to students in 1858. He found the rigor he sought without much difficulty, but did not bother to publish what he regarded as mere common sense until 1872, when he wished to publish something in honor of his father. In his book *Stetigkeit und irrationale Zahlen* (*Continuity and Irrational Numbers*) he referred to Newton's definition of a real number:

... the way in which the irrational numbers are usually introduced is based directly upon the conception of extensive magnitudes—which itself is nowhere carefully defined—and explains number as the result of measuring such a magnitude by another of the same kind. Instead of this I demand that arithmetic shall be developed out of itself.

As Dedekind saw the matter, it was really the *totality* of rational numbers that defined a ratio of continuous magnitudes. Although one might not be able to say that two continuous quantities a and b had a ratio *equal* to, or defined by, a ratio $m : n$ of two integers, an inequality such as $ma < nb$ could be interpreted as saying that the real number $a : b$ (whatever it was) was *less than* the rational number n/m . Thus a positive real number could be defined as a way of dividing the positive rational numbers into two classes, those that were larger than the number and those that were equal to it or smaller, and every member of the first class was larger than every member of the second class. But, so reasoned Dedekind, once the positive rational numbers have been partitioned in this way, the two classes themselves can be regarded as the number. They are a well-defined object, and one

can define arithmetic operations on such classes so that the resulting system has all the properties we want the real numbers to have, especially the essential one for calculus: continuity. Dedekind claimed that in this way he was able to prove rigorously for the first time that $\sqrt{2}\sqrt{3} = \sqrt{6}$.¹⁴

The practical-minded reader who is content to use approximations will probably be getting somewhat impatient with the discussion at this point and asking if it was really necessary to go to so much trouble to satisfy a pedantic desire for rigor. Such a reader will be in good company. Many prominent mathematicians of the time asked precisely that question. One of them was Rudolf Lipschitz (1832–1903). Lipschitz didn't see what the fuss was about, and he objected to Dedekind's claims of originality (Scharlau, 1986, p. 58). In 1876 he wrote to Dedekind:

I do not deny the validity of your definition, but I am nevertheless of the opinion that it differs only in form, not in substance, from what was done by the ancients. I can only say that I consider the definition given by Euclid... to be just as satisfactory as your definition. For that reason, I wish you would drop the claim that such propositions as $\sqrt{2}\sqrt{3} = \sqrt{6}$ have never been proved. I think the French readers especially will share my conviction that Euclid's book provided necessary and sufficient grounds for proving these things.

Dedekind refused to back down. He replied (Scharlau, 1986, pp. 64–65):

I have never imagined that my concept of the irrational numbers has any particular merit; otherwise I should not have kept it to myself for nearly fourteen years. Quite the reverse, I have always been convinced that any well-educated mathematician who seriously set himself the task of developing this subject rigorously would be bound to succeed... Do you really believe that such a proof can be found in any book? I have searched through a large collection of works from many countries on this point, and what does one find? Nothing but the crudest circular reasoning, to the effect that $\sqrt{a}\sqrt{b} = \sqrt{ab}$ because $(\sqrt{a}\sqrt{b})^2 = (\sqrt{a})^2(\sqrt{b})^2 = ab$; not the slightest explanation of how to multiply two irrational numbers. The proposition $(mn)^2 = m^2n^2$, which is proved for rational numbers, is used unthinkingly for irrational numbers. Is it not scandalous that the teaching of mathematics in schools is regarded as a particularly good means to develop the power of reasoning, while no other discipline (for example, grammar) would tolerate such gross offenses against logic for a minute? If one is to proceed scientifically, or cannot do so for lack of time, one should at least honestly tell the pupil to believe a proposition on the word of the teacher, which the students are willing to do anyway. That is better than destroying the pure, noble instinct for correct proofs by giving spurious ones.

¹⁴ In his paper (1992) David Fowler (1937–2004) investigated a number of approaches to the arithmetization of the real numbers and showed how the specific equation $\sqrt{2}\sqrt{3} = \sqrt{6}$ could have been proved geometrically, and also how difficult this proof would have been using many other natural approaches.

Mathematicians have accepted the need for Dedekind's rigor in the teaching of mathematics majors, although the idea of defining real numbers as partitions of the rational numbers (Dedekind cuts) is no longer the most popular approach to that rigor. More often, students are now given a set of axioms for the real numbers and asked to accept on faith that those axioms are consistent and that they characterize a set that has the properties of a geometric line. Only a few books attempt to start with the rational numbers and construct the real numbers. Those that do tend to follow an alternative approach, defining a real number to be a sequence of rational numbers (more precisely, an equivalence class of such sequences, one of which is the sequence of successive decimal approximations to the number).

2.3. Imaginary and complex numbers. Although imaginary numbers seem more abstract to moderns than irrational numbers, that is because their physical interpretation is more remote from everyday experience. One interpretation of $i = \sqrt{-1}$, for example, is as a rotation through a right angle (the effect of multiplying by i in the complex plane). We have an intuitive concept of the length of a line segment and decimal approximations to describe that length as a number; that is what gives us confidence that irrational numbers really are numbers. But it is difficult to think of a rotation as a number. On the other hand, the rules for multiplying complex numbers—at least those whose real and imaginary parts are rational—are much simpler and easier to understand than the definition just given for irrationals. In fact, complex numbers were understood before real numbers were properly defined; mathematicians began trying to make sense of them as soon as there was a clear need to do so. That need came not, as one might expect, from trying to solve quadratic equations such as $x^2 - 2x + 2 = 0$, where the quadratic formula produces $x = -1 \pm \sqrt{-1}$. It was possible in this case simply to say that the equation had no solution. On the other hand, as discussed in Chapter 14, the sixteenth-century Italian mathematicians succeeded in giving an arithmetic solution of the general cubic equation. However, the algorithm for finding the solution had the peculiar property that it involved taking the square root of a negative number precisely when there were three real solutions. Looking at their algorithm as a formula, one would find that the solution of the equation $x^3 - 7x + 6 = 0$ is

$$x = \sqrt[3]{3 - \sqrt{-\frac{100}{27}}} - \sqrt[3]{3 + \sqrt{-\frac{100}{27}}}.$$

We cannot say that the equation has no roots, since it obviously has 1, 2, and -3 as roots. Thus the challenge arose: Make sense of this formula. Make it say "1, 2, and -3 ."

This challenge was taken up by Rafael Bombelli (1526-1572), an engineer in the service of an Italian nobleman. Bombelli was the author of a treatise on algebra which he wrote in 1560, but which was not published until 1572. In that treatise he invented the name "plus of minus" to denote a square root of -1 and "minus of minus" for its negative. He did not think of these two concepts as different numbers, but rather as the *same* number being added in the first case and subtracted in the second. What is most important is that he realized what rules must apply to them in computation: plus of minus times plus of minus makes minus and minus of minus times minus of minus makes minus, while plus of minus times minus of minus makes plus. Bombelli had no systematic way of taking the cube root of a complex number. In considering the equation $x^3 = 15x + 4$, he found by applying the formula that

$x = \sqrt[3]{2 + \sqrt{-121}} + \sqrt[3]{2 - \sqrt{-121}}$. In this case, however, Bombelli was able to work backward, since he knew in advance that one root is 4; the problem was to make the formula say "4." Bombelli had the idea that the two cube roots must consist of real numbers together with his "plus of minus" or "minus of minus." Since the imaginary parts in the sum of the two cube roots must cancel out and the real parts must add up to 4, it seems obvious that the real parts of the cube roots must be 2. In our terms, the cube roots must be $2 \pm t\sqrt{-1}$ for some t . Then since the cube of the cube roots must be $2 \pm 11\sqrt{-1}$ (what Bombelli called 2 plus 11 times "plus of minus"), it is clear that the cube roots must be 2 plus "plus of minus" and 2 minus "plus of minus," that is, $2 \pm \sqrt{-1}$. As a way of solving the equation, this reasoning is circular, but it does allow the formula for solving the cubic equation to make sense.

In an attempt to make these numbers more familiar, the English mathematician John Wallis (1616–1703) pointed out that while no positive or negative number could have a negative square, nevertheless it is also true that no physical quantity can be negative, that is, less than nothing. Yet negative numbers were accepted and interpreted as retreats when the numbers measure advances along a line. Wallis thought that what was allowed in lines might also apply to planes, pointing out that if 30 acres are reclaimed from the sea, and 40 acres are flooded, the net amount "gained" from the sea would be -10 acres. Although he did not say so, it appears that he regarded this real loss of 10 acres as an imaginary gain of a square of land $\sqrt{-435600} = 660\sqrt{-1}$ feet on a side.

What he did say in his 1673 treatise on algebra was that one could represent $\sqrt{-bc}$ as the mean proportional between $-b$ and c . The mean proportional is easily found for two positive line segments b and c . Simply lay them end to end, use the union as the diameter of a circle, and draw the half-chord perpendicular to that diameter at the point where the two segments meet. That half-chord is the mean proportional. When one of the numbers ($-b$) was regarded as negative, Wallis regarded the negative quantity as an oppositely directed line segment. He then modified the construction of the mean proportional between the two segments. When two oppositely directed line segments are joined end to end, one end of the shorter segment lies between the point where the two segments meet and the other end of the longer segment, so that the point where the segments meet lies *outside* the circle passing through the other two endpoints. Wallis interpreted the mean proportional as the tangent to the circle from the point where the two segments meet. Thus, whereas the mean proportional between two positive quantities is represented as a sine, that between a positive and negative quantity is represented as a tangent.

Wallis applied this procedure in an "imaginary" construction problem. First he stated the following "real" problem. Given a triangle having side AP of length 20, side PB of length 15, and altitude PC of length 12, find the length of side AB , taken as base in Fig. 2. Wallis pointed out that two solutions were possible. Using the foot of the altitude as the reference point C and applying the Pythagorean theorem twice, he found that the possible lengths of AB were 16 ± 9 , that is, 7 and 25. This construction is a well-known method of solving quadratic equations geometrically, given earlier by Descartes. It always works when the roots are real, whether positive or negative. He then proposed reversing the data, in effect considering an impossible triangle having side AP of length 20, side PB of length 12, and altitude PC of length 15. Although the algebraic problem has no real solution, a fact verified by

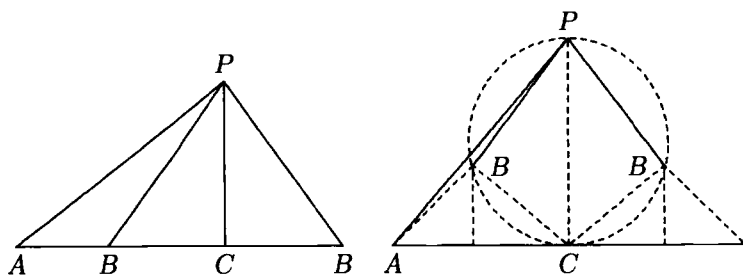


FIGURE 2. Wallis' geometric solution of quadratic equation with real roots (left) and complex roots (right).

the geometric figure (Fig. 2), one could certainly draw the two line segments AB . These line segments could therefore be interpreted as the numerical solutions of the equation, representing a triangle with one side having imaginary length.

The rules given by Bombelli made imaginary and complex numbers accessible, and they turned out to be very convenient in many formulas. Euler made free use of them, studying power series in which the variables were allowed to be complex numbers and deriving a famous formula

$$e^{v\sqrt{-1}} = \cos v + \sqrt{-1} \sin v.$$

Euler derived this result in a paper on ballistics written around 1727 (see Smith, 1929, pp. 95–98), just after he moved to Russia. But he had no thought of representing $\sqrt{-1}$ as we now do, on a line perpendicular to the real axis.

Wallis' work had given the first indication that complex numbers would have to be interpreted as line segments in a plane, a discovery made again a century later by the Norwegian surveyor Caspar Wessel (1745–1818). The only mathematical paper he ever wrote was delivered to the Royal Academy in Copenhagen, Denmark in 1797, but he had been in possession of the results for about a decade at that time. In that paper (Smith, 1929, pp. 55–66), he explained how to multiply lines in a plane by multiplying their lengths and adding the angles they make with a given reference line, on which a length is chosen to represent +1:

Let +1 designate the positive rectilinear unit and $+\epsilon$ a certain other unit perpendicular to the positive unit and having the same origin; the direction angle of +1 will be equal to 0° , that of -1 to 180° , that of $+\epsilon$ to 90° , and that of $-\epsilon$ to -90° or 270° . By the rule that the direction angle of the product shall equal the sum of the angles of the factors, we have: $(+1)(+1) = +1$; $(+1)(-1) = -1$; $(-1)(-1) = +1$; $(+1)(+\epsilon) = +\epsilon$; $(+1)(-\epsilon) = -\epsilon$; $(-1)(+\epsilon) = -\epsilon$; $(-1)(-\epsilon) = +\epsilon$; $(+\epsilon)(+\epsilon) = -1$; $(+\epsilon)(-\epsilon) = +1$; $(-\epsilon)(-\epsilon) = -1$. From this it is seen that ϵ is equal to $\sqrt{-1}$. [Smith, 1929, p. 60]

Wessel noticed the connection of these rules with the addition and subtraction formulas for sign and cosine and gave the formula $(\cos x + \epsilon \sin x)(\cos y + \epsilon \sin y) = \cos(x + y) + \epsilon \sin(x + y)$. On that basis he was able to reduce the extraction of the n th root of a complex number to extracting the same root for a positive real number and dividing the polar angle by n .

The reaction of the mathematical community to this simple but profound idea was less than overwhelming. Wessel's work was forgotten for a full century. In the meantime another mathematician by avocation, the French accountant Jean Argand (1768–1822), published the small book *Essai sur une manière de représenter les quantités imaginaires dans les constructions géométriques* at his own expense in 1806, modestly omitting to name himself as its author, in which he advocated essentially the same idea, thinking, as Wallis had done, of an imaginary number as the mean proportional between a positive number and a negative number. Through a complicated series of events this book and its author gradually became known in the mathematical community. There was, however, resistance to the idea of interpreting complex numbers geometrically, since they had arisen in algebra. But geometry was essential to the algebra of complex numbers, as shown by the fact that a proof of the fundamental theorem of algebra by Gauss in 1799 is based on the idea of intersecting curves in a plane. The lemmas that Gauss used for the proof had been proved earlier by Euler using the algebra of imaginary numbers, but Gauss gave a new proof using only real numbers, precisely to avoid invoking any properties of imaginary numbers.

Even though he avoided the algebra of imaginary numbers, Gauss still needed the continuity properties of real numbers, which, as we just saw, were not fully arithmetized until many years later.¹⁵ Continuity was a geometric property not explicitly found in Euclid, but Gauss expressed the opinion that continuity could be arithmetized. In giving a fifth proof of this theorem half a century later, he made full use of complex numbers. In fact, the complex plane is sometimes called the *Gaussian plane*.

2.4. Infinite numbers. The problem of infinity has occupied mathematicians for a very long time. Neither arithmetic nor geometry can place any preassigned limit on the sizes of objects. An integer can be as large as we like, and a line can be bisected as many times as we like. These are *potential infinities* and *potential infinitesimals*. Geometry can lead to the concept of an *actual infinity* and an *actual infinitesimal*. A line, plane, or solid is an infinite set of points; and in a sense a point is an infinitesimal (infinitely short) line, a line is an infinitesimal (infinitely narrow) plane, and a plane is an infinitesimal (infinitely thin) solid. These notions of the infinite and the infinitesimal present a logical problem for beings whose experience extends over only a finite amount of space, whose senses cannot resolve impressions below a certain threshold, and whose reasoning is presented using a finite set of words. The difficulties of dividing by zero and the problem of incommensurables, mentioned above, are two manifestations of this difficulty. We shall see others in later chapters.

The infinite in Hindu mathematics. Early Hindu mathematics had a prominent metaphysical component that manifested itself in the handling of the infinite. The Hindus accepted an actual infinity and classified different kinds of infinities. This part of Hindu mathematics is particularly noticeable with the Jains. They classified numbers as enumerable, unenumerable, and infinite, and space as one-dimensional, two-dimensional, three-dimensional, and infinitely infinite. Further, they seem to have given a classification of infinite numbers remarkably similar to

¹⁵ The Czech scholar Bernard Bolzano (1781–1848) showed how to approach the idea of continuity analytically in an 1817 paper. One could argue that his work anticipated Dedekind's arithmetization of real numbers.

the modern theory of infinite ordinals. The idea is to progress through the finite numbers $2, 3, 4, \dots$ until the “first unenumerable” number is reached. This number corresponds to what is now called ω , the first infinite ordinal number. Then, exactly as in modern set theory, one can consider the unenumerable numbers $\omega + 1$, $\omega + 2, \dots, \omega^2$, and so on. We do not have enough specifics to say any more, but there is a very strong temptation to say that the Jaina classification of enumerable, unenumerable, infinite corresponds to our modern classification of finite, countably infinite, and uncountably infinite, but of course it is only a coincidental prefiguration.

Infinite ordinals and cardinals. A fuller account of the creation of the theory of cardinal and ordinal numbers in connection with set theory is given in Chapter 19. At this point, we merely note that these theories were created along with set theory in the late nineteenth century through the work of several mathematicians, most prominently Georg Cantor (1854–1918). The relation between cardinal and ordinal numbers is an important one that has led to a large amount of research. Briefly, ordinal numbers arise from continuing the ordinary series of natural numbers “past infinity.” Cardinal numbers arise from comparing two sets by matching their elements in a one-to-one manner.

3. Combinatorics

From earliest times mathematicians have been concerned with counting things and with space, that is, with the arrangement of objects of interest. Counting arrangements of things became a separate area of study within mathematics. We now call this area *combinatorics*, and it has ramified to include a number of distinct areas of interest, such as formulas for summation of powers, graph theory, magic squares, Latin squares, Room squares, and others. We have seen already that magic squares were used in divination, and there is a very prominent connection between this area and some varieties of mystical thinking. It may be coincidental that the elementary parts of probability theory, the parts that students find most frustrating, involve these sophisticated methods of counting. Probability is the mathematization of possible outcomes of events, exactly the matters that are of interest to people who consult oracles. These hypothetical happenings are usually too many to list, and some systematic way of counting them is needed.

3.1. Summation rules. The earliest example of a summation problem comes from the Ahmose Papyrus. Problem 79 describes seven houses in which there are seven cats, each of which had eaten seven mice, each of which had eaten seven seeds, each of which would have produced seven *hekats* of grain if sown. The author asks for the total, that is, for the sum $7 + 7^2 + 7^3 + 7^4 + 7^5$, and gives the answer correctly as 19,607. The same summation with a different illustration is found in Fibonacci's *Liber abaci* of 1202. In this example we encounter the summation of a finite geometric progression.

A similar example is Problem 34 of Chapter 3 of the *Sun Zi Suan Jing*, which tells of 9 hillsides, on each of which 9 trees are growing, with 9 branches on each tree, 9 bird's nests on each branch, and 9 birds in each nest. Each bird has 9 young, each young bird has 9 feathers, and each feather has 9 colors. The problem asks for the total number of each kind of object and gives the answer: 81 trees, 729 branches, 6561 nests, 59,049 birds, 531,441 young birds, 4,782,969 feathers, and

43,046,721 colors. It does *not* ask for the sum of this series, which indeed is an absurd operation, given that the objects are of different kinds.

Hindu mathematicians gave rules for summing geometric progressions and also the terms of arithmetic progressions and their squares. In Section 3 of Chapter 12 of the *Brahmasphutasiddhanta* (Colebrooke, 1817, pp. 290–295), Brahmagupta gives four rules for dealing with arithmetic progressions. The first rule gives the sum of an arithmetic progression as its average value (half the sum of its first and last terms) times its period, which is the number of terms in the progression. We would write this rule as the formula

$$\sum_{k=0}^n (a + kd) = (n + 1) \frac{a + (a + nd)}{2} = (n + 1)a + d \frac{n(n + 1)}{2}.$$

For the case $a = 0$ and $d = 1$, this formula gives the familiar rule

$$\sum_{k=1}^n k = \frac{n(n + 1)}{2},$$

and Brahmagupta then says that the sum of the squares will be this number multiplied by twice the period added to 1 and divided by 3; in other words

$$\sum_{k=1}^n k^2 = \frac{n(n + 1)(2n + 1)}{6}.$$

For visual proofs of these results Brahmagupta recommended using piles of balls or cubes.

These same rules were given in Chapter 5 of Bhaskara's *Lilavati* (Colebrooke, 1817, pp. 51–57). Bhaskara goes a step further, saying, "The sum of the cubes of the numbers one, and so forth, is pronounced by the ancients equal to the square of the addition." This also is the correct rule that we write as

$$\sum_{k=1}^n k^3 = \left(\sum_{k=1}^n k \right)^2 = \left(\frac{n(n + 1)}{2} \right)^2.$$

Bhaskara also gives the general rule for the sum of a geometric progression of $n + 1$ terms $(a, ar, ar^2, \dots, ar^n)$ in terms that amount to $a(r^{n+1} - 1)/(r - 1)$. He illustrates this rule with several examples, finding that $2 + 6 + 18 + 54 + 162 + 486 + 1456 = 2(3^7 - 1)/2 = 3^7 - 1 = 2186$.

About a century later than Bhaskara, Fibonacci's *Liber quadratorum* gives the same rule for the sum of the squares (Proposition 10): "If beginning with the unity, a number of consecutive numbers, both even and odd numbers, are taken in order, then the triple product of the last number and the number following it and the sum of the two is equal to six times the sum of the squares of all the numbers, namely from the unity to the last."

Proposition 11 gives a more elaborate summation rule, which we can express simply as

$$(2n + 1)(2n + 3)(4n + 4) = 12(1^2 + 3^2 + 5^2 + \dots + (2n + 1)^2).$$

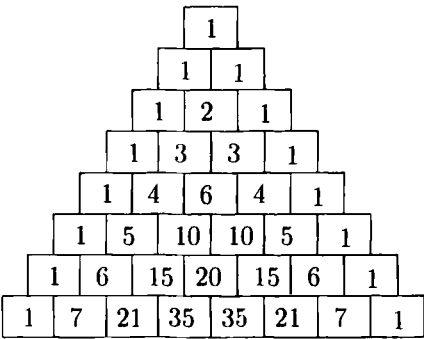


FIGURE 3. The *Meru Prastara*.

Permutations and combinations. The metaphysics of the Jainas, based on a classification of sentient beings according to the number of senses possessed, led them to a mathematical topic related to number theory. They called it *vikalpa*, and we know it as the basic part of combinatorics.

A typical question might be: “How many groups of three can be formed from a collection of five objects?” We know the answer, as did the early Jaina mathematicians. In the *Bhagabati Sutra*, written about 300 BCE, the author asks how many philosophical systems can be formed by taking a certain number of doctrines from a given list of basic doctrines. After giving the answers for 2, 3, 4, and so on, the author says that enumerable, unenumerable, and infinite numbers of things can be discussed, and “as the number of combinations are formed, all of them must be worked out.”

The general process for computing combinatorial coefficients was known to the Hindus at an early date. Combinatorial questions seemed to arise everywhere for the Hindus, not only in the examples just given but also in a work on medicine dating from the sixth century BCE (Biggs, 1979, p. 114) that poses the problem of the number of different flavors that can be made by choosing subsets of six basic flavors (bitter, sour, salty, astringent, sweet, hot). The author gives the answer as $6 + 15 + 20 + 15 + 6 + 1$, that is, 63. We recognize here the combinatorial coefficients that give the subsets of various sizes that can be formed from six elements. The author did not count the possibility of no flavor at all.

Combinatorics also arose in the study of Sanskrit in the third century BCE when the writer Pingala gave a rule for finding the number of different words that could be formed from a given number of letters. This rule was written very obscurely, but a commentator named Halayudha, who is believed to have lived in the tenth century CE (Needham, 1959, p. 37), explained it as follows. First draw a square. Below it and starting from the middle of the lower side, draw two squares. Then draw three squares below these, and so on. Write the number 1 in the middle of the top square and inside the first and last squares of each row. Inside every other square the number to be written is the sum of the numbers in the two squares above it and overlapping it.

This array of numbers, which is known as *Pascal's triangle* because of a treatise on it written by Pascal in the seventeenth century, was studied four centuries before Pascal by Jordanus Nemorarius, who developed many of its properties (Hughes, 1989). Pascal's triangle also occurs in Chinese manuscripts some four centuries before Pascal's treatise. In China the inspiration for the study of this diagram arose in connection with the extraction of cube roots and higher roots. The diagram appears in the *Xiangjie Jiuzhang Suan Fa* (*Detailed Analysis of the Mathematical Methods in the Nine Chapters*) of Yang Hui, written in 1261 (Li and Du, 1987, p. 122). But in India we find it 300 years before it was published in China and 700 years before Pascal. Moreover it purports to be only a clarification of a rule invented 1200 years earlier!¹⁶ Its Sanskrit name is *Meru Prastara* (see Fig. 3), which means the *staircase of Mount Meru*.¹⁷

According to Singh (1985), Pingala's work on poetry also leads to another interesting combinatorial topic, recognized as such by the Hindu mathematicians. We treated this topic above as number theory, but it will bear repeating as combinatorics. To simplify the explanation as much as possible, suppose that a line of poetry is to be written using short beats and long ones, a long one being equivalent to two short ones. If a line contains n beats, how many arrangements are possible. Just to get started, we see that there is obviously one line of one beat (short), two lines of two beats (two short or one long), three lines of three beats (short-long, long-short, short-short-short), and five lines of four beats (long-long, short-short-long, short-long-short, long-short-short, short-short-short-short). Since a line with $n + 1$ beats must begin with either a short or a long beat, we observe that those beginning with a short beat are in one-to-one correspondence with the lines of n beats, all of which can be obtained by removing the initial short beat, while those beginning with a long beat are in a similar correspondence with lines of $n - 1$ beats. It follows that the number of lines with $n + 1$ beats is the sum of the numbers with $n - 1$ and n . Once again we generate the Fibonacci sequence.

Bhaskara II knew the rules for combinatorial coefficients very well. In Chapter 4 of the *Lilavati* (Colebrooke, 1817, pp. 49–50), he gives an example of hexameter and asks how many possible combinations of long and short syllables are possible. He prescribes setting the numbers from 1 to 6 down “in direct and inverse order,” that is, setting down the 2×6 matrix

$$\begin{array}{cccccc} 6 & 5 & 4 & 3 & 2 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{array}$$

From this array, by forming the products from the left and dividing, he finds the number of verses with different numbers of short syllables from 1 to 5 as

$$\frac{6}{1} = 6, \quad \frac{6 \cdot 5}{1 \cdot 2} = 15, \quad \frac{6 \cdot 5 \cdot 4}{1 \cdot 2 \cdot 3} = 20, \quad \frac{6 \cdot 5 \cdot 4 \cdot 3}{1 \cdot 2 \cdot 3 \cdot 4} = 15, \quad \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 6.$$

¹⁶ That claim cannot be verified, however. Evidence indicates that knowledge of the combinatorial coefficients arose in India around the time of Aryabhata I, in the sixth century (Biggs, 1979, p. 115).

¹⁷ In Hindu mythology Mount Meru plays a role similar to that of Mount Olympus in Greek mythology. One Sanskrit dictionary gives this mathematical meaning of *Meru Prastara* as a separate entry. The word *prastara* apparently has some relation to the notion of expansion as used in connection with the binomial theorem.

Bhaskara recognized that there was one more possibility, six short syllables, but did not mention the possibility of no short syllables. The convention that an empty product equals 1 was not part of his experience.

Magic squares and Latin squares. In 1274 (see Li and Du, 1987, p. 166) Yang Hui wrote *Xugu Zhaiqi Suanfa* (*Continuation of Ancient Mathematical Methods for Elucidating the Strange* [Properties of Numbers]), in which he listed magic squares of order up to 10. According to Biggs (1979, p. 121), there is evidence that the topic of magic squares had reached a high degree of development in China before this time and that Yang Hui was merely listing ancient results that were not a topic of current research, since he seems to have no concept of a general rule for constructing magic squares. Magic-square-type figures were a source of fascination in Korea, and they also seem to have spread west from China. Whether from China or as an indigenous product, magic squares of order up to 6 appear in the Muslim world around the year 990 (Biggs, 1979, p. 119) and squares of order up to 9 are mentioned. Rules for constructing such squares were given by the Muslim scholar al-Buni (d. 1225). From the Muslim world, they entered Europe around the year 1315 in the works of the Greek scholar named Manuel Moschopoulos, who was claimed as a student by Maximus Planudes, who was mentioned in Chapter 2. They exerted a fascination on European scholars also, and the artist Albrecht Dürer incorporated a 4×4 magic square in his famous engraving *Melencolia*, with the year of its composition (1514) in the bottom row. The most fascinating thing about them is their sheer number. Difficult as they are to construct, there are nevertheless 880 distinct 4×4 magic squares.

Magic squares occur profusely throughout Indian, Chinese, and Japanese mathematics, alongside more elaborate numerical figures such as magic circles and magic hexagons. A variant of the idea of a magic square is that of a Latin square, an $n \times n$ array in which each of n letters appears once in each row and once in each column. It is easy to construct such a square by writing the letters down in order along the first row and then cyclically permuting them by one step in each subsequent row. To make the problem harder, mathematicians beginning with Euler in 1781 have sought pairs of *orthogonal* Latin squares: that is, two Latin squares that can be superimposed in such a way that each of the n^2 possible ordered pairs of letters occurs exactly once. An example, given by Biggs (1979, p. 123), with one of the squares using Greek letters for additional clarity, is

$$\begin{array}{cccc} A\alpha & B\beta & C\gamma & D\delta \\ B\gamma & A\delta & D\alpha & C\beta \\ C\delta & D\gamma & A\beta & B\alpha \\ D\beta & C\alpha & B\delta & A\gamma \end{array}$$

Modern combinatorics. The usefulness of combinatorics in elementary probability has already been noted. It is an interesting exercise to compute, for example, the probability of a particular poker hand, say a full house, and see why the rules of the game make three of a kind a better hand (because less likely) than two pairs.

The strongest impetus to combinatorial studies in Europe, however, came from Gottfried Wilhelm Leibniz (1646–1716), who is best remembered for his brilliant discoveries in the calculus. He was also a profound philosopher and a diplomat with a deep interest in Oriental cultures. Many fundamental results are found in his *De arte combinatoria*, published in 1666. In this work Leibniz gave tables of



Albrecht Dürer's *Melencolia*, containing a magic square showing the date of composition as 1514. © Corbis Images (No. BE005826).

the number of permutations of n objects. There are many very curious aspects of this work. Although written mostly in Latin, it is rather polyglot. Leibniz occasionally breaks into Greek or German, and the tables are labeled with Hebrew letters. For permutations Leibniz used the word *numerus* to denote the size of the set from which objects are chosen, and *exponent* (literally, *placing out*) for the

number of objects chosen. The total number of permutations of a number of objects he called its *variationes*, and for the number of combinations of a set of objects taken, say, four at a time, he wrote *con4natio*, an abbreviation for *conquattuornatio*. The case of two objects taken at a time provides the modern word *combination*. These combinations, now called *binomial coefficients*, were referred to generically as *complexiones*. The first problem posed by Leibniz was: *Given the numerus and the exponent, find the complexiones*. In other words, given n and k , find the number of combinations of n things taken k at a time.

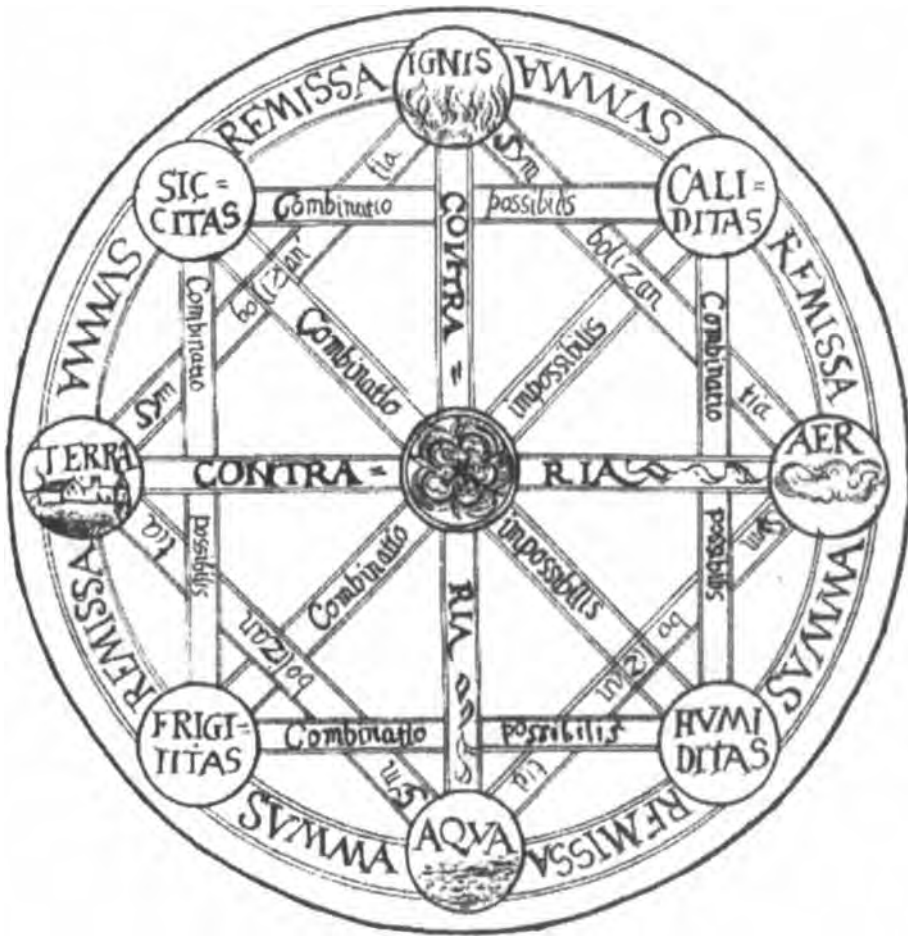
Like the Hindu mathematicians, Leibniz applied combinatorics to poetry and music. He considered the hexameter lines possible with the Guido scale *ut, re, mi, fa, sol, la*, finding a total of 187,920.¹⁸

De arte combinatoria contains 12 sophisticated counting problems and a number of exotic applications of the counting techniques. It appears that Leibniz intended these techniques to be a source by which all possible propositions about the world could be generated. Then, combined with a good logic checker, this technique would provide the key to all knowledge. His intent was philosophical as well as mathematical, as evidenced by his claimed mathematical proof of the existence of God at the beginning of the work. Thus once again, this particular area of mathematics seems to be linked, more than other kinds of mathematics, with mysticism. The frontispiece of *De arte combinatoria* shows a mystical arrangement of the opposite pairs wet/dry, cold/hot, with the four elements of earth, air, fire, and water as cardinal points. This figure resembles an elaborate version of the famous ying/yang symbol from Chinese philosophy and it also recalls the proposition generator of the mystic theologian Ramon Lull (1232–1316), which consisted of a series of nested circles with words inscribed on them. When rotated independently, they would generate sentences. Leibniz was familiar with Lull's work, but was not a proponent of it.

The seeds planted by Leibniz in *De arte combinatoria* sprouted and grew during the nineteenth century, as problems from algebra, probability, and topology required sophisticated techniques of counting. One of the pioneers was the British clergyman Thomas Kirkman (1806–1895). The first combinatorial problem he worked on was posed in the *Lady's and Gentleman's Diary* in 1844: *Determine the maximum number of distinct sets of p symbols that can be formed from a set of n symbols subject to the restriction that no combination of q symbols can be repeated in different sets*. Kirkman himself posed a related problem in the same journal five years later: *Fifteen young ladies in a school walk out three abreast for 7 days in succession; it is required to arrange them daily so that no two shall walk twice abreast*. This problem is an early example of a problem in combinatorial design. The problem of constructing a Latin square is another example. This kind of combinatorial design has a practical application in the scheduling of athletic tournaments, and in fact colleagues of the author specializing in combinatorial design procured a contract to design the schedule for the short-lived XFL Football League in 2001.

We shall terminate our discussion of combinatorics with these nineteenth-century results. We note in parting that it remains an area with a plenitude of unsolved problems whose statement can be understood without long preparation.

¹⁸ The first five of these tones are the first syllables of a medieval Latin chant on ascending tones. The replacement of *ut* by the modern *do* came later.



Frontispiece of *De arte combinatoria*, from Vol. 5, p. 7 of the collected works of Leibniz. © Georg Olms Verlag, Hildesheim.

It was for that reason that combinatorial problems were used as the mathematical background of the film *Good Will Hunting*.

Questions and problems

8.1. We know a mathematical algorithm for computing as many decimal digits of $\sqrt{2}$ as we have time for, and $\sqrt{2}$ has a precise representation in Euclidean geometry as the ratio of the diagonal of a square to its side. It is a provable theorem of Euclidean geometry that that ratio is the same for all squares, so that two observers using different squares should get the same result. To the extent that physical space really is Euclidean, this definition makes it possible to determine $\sqrt{2}$ empirically by measuring the sides and diagonals of physical squares. In that sense, we could theoretically determine $\sqrt{2}$ with arbitrarily prescribed precision by physical measurements. In particular, it makes perfectly good sense to ask what the 50th decimal digit of $\sqrt{2}$ is—it happens to be 4, but rounds up to 5—and we could try to get instruments precise enough to yield this result from measurement.

Consider, in contrast, the case of a physical constant, say the universal gravitational constant, usually denoted G_0 , which occurs in Newton's law of gravitation:

$$F = G_0 \frac{Mm}{r^2}.$$

Here F is the force each of two bodies exerts on the other, M and m are the masses of the two bodies, and r is the distance between their centers of gravity. The accepted value of G_0 , given as upper and lower assured limits, is $6.674215 \pm 0.000092 \text{ N} \cdot \text{m}^2/\text{kg}^2$, although some recent measurements have cast doubt on this value. From a mathematical point of view, G_0 is determined by the equation

$$G_0 = \frac{Fr^2}{Mm},$$

and its value is found—as Cavendish actually did—by putting two known masses M and m at a known distance r from each other and measuring the force each exerts on the other. The assertion that the ratio Fr^2/Mm is the same for all masses and all distances is precisely the content of *Newton's law of gravity*, so that two experimenters using different masses and different distances should get the same result. But Newton's law of gravity is not deducible from axioms; it is, rather, an empirical hypothesis, to be judged by its explanatory power and its consistency with observation. What should we conclude if two experimenters do *not* get the same result for the value of G_0 ? Did one of them do something wrong, or is Newton's law not applicable in all cases? Does it even *make sense* to ask what the 50th decimal digit of G_0 is?

8.2. You can represent \sqrt{ab} geometrically by putting a line of length b end-to-end with a line of length a , drawing a circle having this new line as diameter, and then drawing the perpendicular to the circle from the point where the two lines meet. To get \sqrt{a} and \sqrt{b} , you would have to use Descartes' unit length I as one of the factors. Is it possible to prove by use of this construction that $\sqrt{ab}I = \sqrt{a}\sqrt{b}$? Was Dedekind justified in claiming that this identity had never been proved?

8.3. Try to give a definition of real numbers—perhaps using decimal expansions—that will enable you to say what the numbers $\sqrt{2}$, $\sqrt{3}$, and $\sqrt{6}$ are, and how they can be added and multiplied. Does your definition enable you to prove that $\sqrt{2}\sqrt{3} = \sqrt{6}$?

8.4. Use the method of infinite descent to prove that $\sqrt{3}$ is irrational. [*Hint:* Assuming that $m^2 = 3n^2$, where m and n are positive integers having no common factor, that is, they are as small as possible, verify that $(m - 3n)^2 = 3(m - n)^2$. Note that $m < 2n$ and hence $m - n < n$, contradicting the minimality of the original m and n .]

8.5. Show that $\sqrt[3]{3}$ is irrational by assuming that $m^3 = 3n^3$ with m and n positive integers having no common factor. [*Hint:* Show that $(m - n)(m^2 + mn + n^2) = 2n^3$. Hence, if p is a prime factor of n , then p divides either $m - n$ or $m^2 + mn + n^2$. In either case p must divide m . Since m and n have no common factor, it follows that $n = 1$.]

8.6. Suppose that x , y , and z are positive integers, no two of which have a common factor, none of which is divisible by 3, and such that $x^3 + y^3 = z^3$. Show that there exist integers p , q , and r such that $z - x = p^3$, $z - y = q^3$, and $x + y = r^3$. Then,

letting $m = r^3 - (p^3 + q^3)$ and $n = 2pqr$, verify from the original equation that $m^3 = 3n^3$, which by Problem 8.5 is impossible if m and n are nonzero. Hence $n = 0$, which means that $p = 0$ or $q = 0$ or $r = 0$, that is, at least one of x and y equals 0. Conclude that no such positive integers x , y , and z can exist.

8.7. Verify that

$$27^5 + 84^5 + 110^5 + 133^5 = 144^5.$$

[See L. J. Lander and T. R. Parkin, "Counterexample to Euler's conjecture on sums of like powers," *Bulletin of the American Mathematical Society*, **72** (1966), p. 1079. Smaller counterexamples to this conjecture have been discovered more recently.]

8.8. Prove Fermat's little theorem by induction on a . [Hint: The theorem can be restated as the assertion that p divides $a^p - a$ for every positive integer a . Use the binomial theorem to show that $(a + 1)^p - (a + 1) = mp + a^p - a$ for some integer m .]

8.9. Verify the law of quadratic reciprocity for the primes 17 and 23 and for 67 and 71.

8.10. Show that the factorization of numbers of the form $m + n\sqrt{-3}$ is *not* unique by finding two different factorizations of 4. Is factorization unique for numbers of the form $m + n\sqrt{-2}$?

8.11. Prove that the number of primes less than or equal to N is at least $\log_2(N/3)$, by proceeding as follows. Let p_1, \dots, p_n be the prime numbers among $1, \dots, N$, and let $\theta(N)$ be the number of square-free integers among $1, \dots, N$, that is, the integers not divisible by any square number. We then have the following relation, since it is known that $\sum_{k=1}^{\infty} (1/k^2) = \pi^2/6$.

$$\begin{aligned} \theta(N) &> N - \sum_{k=1}^n \left[\frac{N}{p_k^2} \right] \\ &> N \left(1 - \sum_{k=1}^n \frac{1}{p_k^2} \right) \\ &> N \left(1 - \sum_{k=2}^{\infty} \frac{1}{k^2} \right) \\ &= N \left(2 - \frac{\pi^2}{6} \right) > \frac{N}{3}. \end{aligned}$$

(Here the square brackets denote the greatest-integer function.) Now a square-free integer k between 1 and N is of the form $k = p_1^{e_1} \cdots p_n^{e_n}$, where each e_j is either 0 or 1. Hence $\theta(N) \leq 2^n$, and so $n > \log_2(N/3)$. This interesting bit of mathematical trivia is due to the Russian-American mathematician Joseph Perott (1854–1924).

8.12. Assuming that $\lim_{n \rightarrow \infty} \frac{\log(n)\pi(n)}{n}$ exists, use Chebyshev's estimates to show that this limit must be 1 and hence that Legendre's estimate cannot be valid beyond the first term.

Part 3

Color Plates

Plate 1. Top to bottom: Problems 49–54 of the Ahmose Papyrus. © The British Museum.



Plate 2. (Overleaf): *Sangaku* on display at the Suguwara Shrine in Mie Prefecture, 1854. Courtesy of Mr. Hidetoshi Fukagawa.



Plate 3. Folio 47 of the Dresden Codex. © Sächsische Landesbibliothek, Dresden.



Plate 4. A branch of a flowering crabapple tree, illustrating the Fibonacci/golden ratio pattern of twig growth.

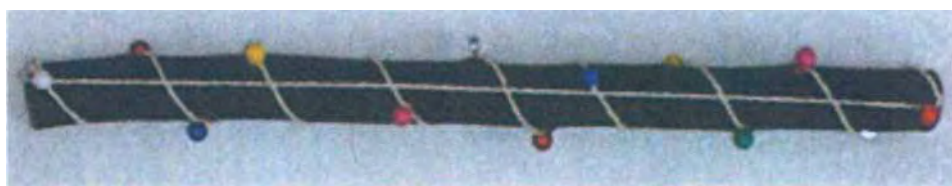


Plate 5. Florence Nightingale's "batwing" or "coxcomb" diagram, forerunner of the modern pie chart. Unlike the modern pie chart, the sectors here do not represent percentages. Rather, they give the monthly death tolls from wounds and disease during the Crimean War of 1854-1855. Nowadays, such data would be presented as a line graph.

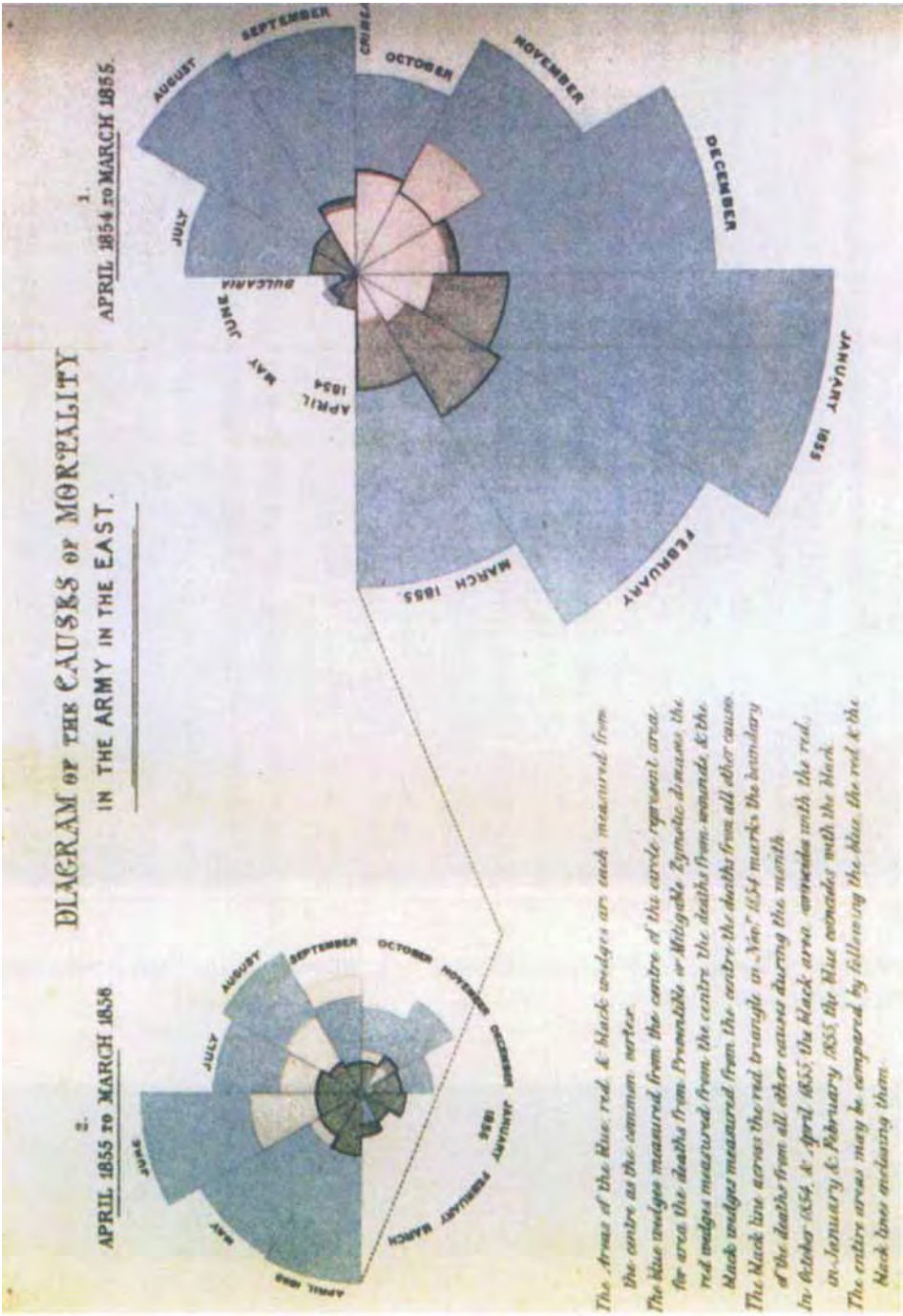


Plate 6. A page of the Archimedes Palimpsest. © Christie's Images Incorporated 2005.



Plate 7. Egyptian field measurers: a wall painting from the tomb of Menna, Scribe of the Fields, around 1400 BCE. © Corbis Images (No. WF004135).

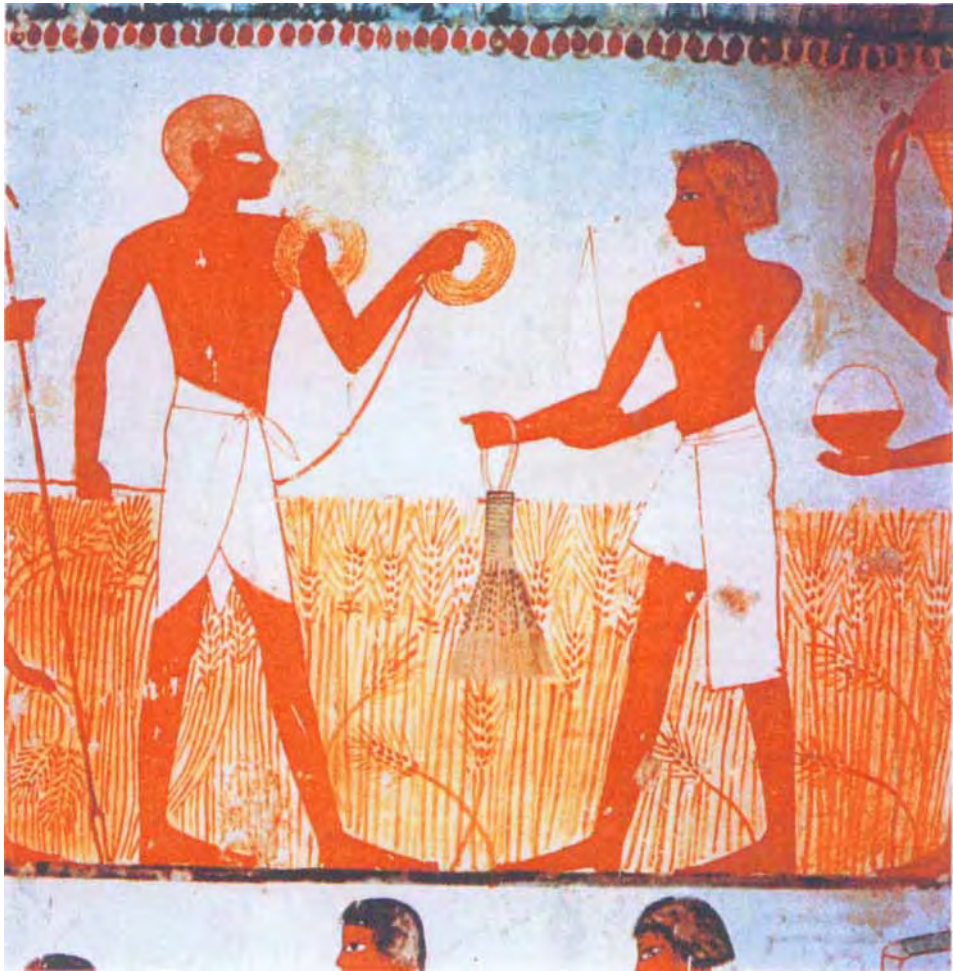


Plate 8. Reconstruction of one of Ptolemy's maps of the world. © The Vatican Library (Urb. gr. 82, ff. 60v-61r).



Plate 9. Portrait of Luca Pacioli in 1495 working with a text of Euclid's *Elements*, by Jacopo de Barbari (1440–1515). It is not certain who the young man behind him is, but it may be Albrecht Dürer. © Corbis Images (No. CS002808).



Part 4

Space

Geometry is a way of organizing our perception of shape. As in arithmetic, we can distinguish levels of sophistication in the development of geometry. The first level is that of measurement: comparing the sizes of objects having different shapes. Measurement is *arithmetic applied to space*, beginning with figures having flat sides or faces. The second level is the study of the proportions among the parts of geometric figures, such as triangles, squares, rectangles, and circles. A good marker for the beginning of this stage of development is the Pythagorean relation for right triangles.

Although regular polygons and polyhedra can be measured using simple dissection techniques, algebra is needed to measure more complicated figures, such as the portion of a disk remaining after three other disks, all tangent to one another and to the original disk and whose radii are in given ratios, are removed from it. One of the uses of geometry in many ancient treatises is as a source of interesting equations to be solved. Although the problems are posed as problems in measurement, the shapes being measured are so unusual that it is hard to think of them as a motive. The suspicion begins to arise that the author's real purpose was to exhibit some algebra.

The Pythagoreans, Plato, and Aristotle gave geometry a unique philosophical grounding that turned it out of the path it would probably have followed otherwise. Their insistence on a logical development based on a system of axioms made explicit many assumptions—especially the parallel postulate—that otherwise might not have been noticed. As a result, there is a marked difference between the mathematical practitioners who learned geometry from Euclid (the medieval Muslims and renaissance Europeans) and those who learned it from a different tradition.

At its highest level, elementary geometry employs algebra and the infinite processes of calculus in order to find the areas and volumes of ever more complicated curvilinear figures. Like arithmetic, geometry has given rise to many specialties, such as projective, analytic, and differential geometry. As more and more general properties of space became mathematized, geometry generated the subject known as *topology*, from Greek roots meaning *theory of position*.

In the four chapters that constitute the present part of our history, we shall look at all these aspects of geometry. In Chapter 9 we study the way space was measured in a number of civilizations. Chapters 10 and 11 form a unit devoted to the most influential form of elementary geometry, the Euclidean geometry that arose in the Hellenistic civilization. Chapter 12 contains a survey of the variety of forms of geometry that have arisen over the past three centuries.

CHAPTER 9

Measurement

The word *geometry* comes from the Greek words *gê*, meaning *earth*, and *metrein*, meaning *measure*. It seems that all human societies have had to measure fields for agriculture or compute the amount of work involved in excavating a building site. That there is a basic similarity in approaches to these problems is attested by the presence of words for circles, rectangles, squares, and triangles in every language. Geometric intuition seems to be innate to human beings. Many different societies independently discovered the Pythagorean theorem, for example. These external similarities conceal certain differences in outlook, however. For example, we are taught to think of a line as having no thickness. But did the Hindus, Egyptians, and others think of it that way? The word *line* comes from the Greek *línon* (Latin *linea*) meaning *string*; the physical object (a stretched string) on which the abstraction is based is clear. Early Hindu work on geometry also uses Sanskrit words for ropes and cords. It is very likely that ancient engineers thought of a line as a physical object: a rope, stretched taut. The quantity of rope was given as a number (length). Geometry at this stage of development was a matter of relating lengths to other quantities of geometric interest, such as areas, slopes, and volumes. It was an application of arithmetic and useful in planning public works projects, for example, since it provided an estimate of the size of a job and hence the number of workers and the amount of materials and time required to excavate and build a structure. It also proved useful in surveying, since geometric relationships could be used to compute inaccessible distances from accessible ones. The fact that similar triangles are the basic tool in this science has caused it to be named *trigonometry*. The elementary rules for measuring regular geometric figures persist in treatises for many centuries. Nearly always the author begins by describing the standard units of length, area, volume, and weight, then presents a variety of procedures that have a great resemblance to the procedures described in all other treatises of the same kind. This geometry, though elementary, should not be thought of as primitive. Textbooks of “practical mathematics” containing exactly this material are still being written and published today.

Although geometry looks very much the same across cultures, there is one place where we must exercise a little care in order to understand it from the point of view of the people we are studying. Textbooks often give approximate values of the number π allegedly used in different cultures without being clear about which constant they mean. When calculating the circumference C of a circle in terms of its diameter, we use the formula $C = \pi d$. When calculating the area of a circle (disk) in terms of its radius, we use the familiar formula $A = \pi r^2$. When calculating the area of a sphere in terms of its radius, we use $A = 4\pi r^2$, and for the volume, the formula $V = \frac{4}{3}\pi r^3$. There are really four different values of π here, depending on the dimension and flatness or curvature. The first formula reflects the fact that the

circumference of a circle is proportional to its diameter. That is, given two circles with circumferences C_i and diameters D_i , $i = 1, 2$, we have $C_1/C_2 = D_1/D_2$. We shall call the ratio C/D the *one-dimensional* π . The second formula implies that if Δ_1 and Δ_2 are the areas of disks of radius r_1 and r_2 and S_1 and S_2 are the areas of squares of sides r_1 and r_2 , then $\Delta_1/\Delta_2 = S_1/S_2$. We shall call the ratio Δ/S the *two-dimensional* π , and similarly for volumes. We won't need the π that occurs in the formula for the area of a sphere, since everybody seemed to relate that one to one of the others. Thus, we are dealing with several direct proportions with different constants of proportionality. It is not obvious that these constants for different dimensions have any simple relationship to one another. That fact requires some digging in geometry to discover. Without the abstract concept of a *constant of proportionality*, when a mathematician is seeking only numerical approximations that accord with observation, there is no reason to suspect any connection between these constants in different dimensions. To be sure, only a small amount of intuition is required to establish the connection, as shown in Problems 9.13 and 9.20, but in any discussion of supposed approximations to π used in different cultures, we need to keep in mind the dimension of the object being studied: Was it a circle, a disk, a cone, a cylinder, a sphere, or a ball?

1. Egypt

Foreigners have been interested in the geometry of the Egyptians for a very long time. In Section 109 of Book 2 of his *History*, the Greek historian Herodotus writes that King Sesostris¹ dug a multitude of canals to carry water to the arid parts of Egypt. He goes on to connect this Egyptian engineering with Greek geometry:

It was also said that this king distributed the land to all the Egyptians, giving an equal quadrilateral farm to each, and that he got his revenue from this, establishing a tax to be paid for it. If the river carried off part of someone's farm, that person would come and let him know what had happened. He would send surveyors to remeasure and determine the amount by which the land had decreased, so that the person would pay less tax in proportion to the loss. It seems likely to me that it was from this source that geometry was found to have come into Greece. For the Greeks learned of the sundial and the twelve parts of the day from the Babylonians.

The main work of Egyptian surveyors was measuring fields. That job corresponds well to the Latin word *agrimensor*, which means *surveyor*. Our word *surveyor* comes through French, but has its origin in the Latin *supervideo*, meaning *I oversee*. The equivalent word in Greek was used by Herodotus in the passage above. He said that the king would send *episkepsoménous kai anametrésontas*, literally *overseeing and remeasuring men*. The process of measuring a field is shown in a painting from the tomb of an Egyptian noble named Menna at Sheikh Abd el-Qurna in Thebes (Plate 7). Menna bore the title Scribe of the Fields of the Lord of the Two Lands during the eighteenth dynasty, probably in the reign of Amenhotep III or Thutmose IV, around 1400 BCE. His job was probably that of

¹ There were several pharaohs with this name. Some authorities believe that the one mentioned by Herodotus was actually Ramses II, who ruled from 1279 to 1212 BCE.

a steward, to oversee planting and harvest. As the painting shows, the instrument used to measure distance was a rope that could be pulled taut. That measuring instrument has given rise to another name often used to refer to these surveyors: *harpedonáptai*, from the words *harpedónē*, meaning *rope*, and *háptein*, meaning *attach*. The philosopher Democritus (d. 357 BCE) boasted, "In demonstration no one ever surpassed me, not even those of the Egyptians called *harpedonáptai*."²

The geometric problems considered in the Egyptian papyri all involve measurement. These problems show considerable insight into the properties of simple geometric figures such as the circle, the triangle, the rectangle, and the pyramid; and they rise to a rather high level of sophistication in computing the area of a hemisphere. Those involving flat boundaries (polygons and pyramids) are correct from the point of view of Euclidean geometry, while those involving curved boundaries (disks and spheres) are correct up to the constant of proportionality chosen.

1.1. Areas. Since the areas of rectangles and triangles are easy to compute, it is understandable that very little attention is given to these problems. Only four problems in the Ahmose Papyrus touch on these questions: Problems 6, 49, 51, and 52 (see Plate 1).

Rectangles, triangles, and trapezoids. Problem 49 involves computing the area of a rectangle that has dimensions 1 *khet* by 10 *khet*s. This in itself would be a trivial problem, except that areas are to be expressed in square cubits rather than square *khet*s. Since a *khet* is 100 cubits, the answer is given correctly as 100,000 square cubits. Problem 51 is a matter of finding the area of a triangle, and it is illustrated by a figure (see Plate 1) showing the triangle. The area is found by multiplying half of the base by the height. In Problem 52, this technique is generalized to a trapezoid, and half of the sum of the upper and lower bases is multiplied by the height.

Of all these problems, the most interesting is Problem 6, which involves a twist that makes it equivalent to a quadratic equation. A rectangle is given having area 12 *cubit strips*; that is, it is equal to an area 1 cubit by 12 cubits, though not of the same shape. The problem is to find its dimensions given that the width is three-fourths of the length ($\frac{2}{3}$ $\frac{4}{5}$ in the notation of the papyrus). The first problem is to "calculate with $\frac{2}{3}$ $\frac{4}{5}$, until 1 is reached," that is, in our language, dividing 1 by $\frac{2}{3}$ $\frac{4}{5}$. The result is $1 \frac{3}{4}$. Then 12 is multiplied by $1 \frac{3}{4}$, yielding 16, after which the scribe takes the *corner* (square root) of 16—unfortunately, without saying how—getting 4 as the length. This is a very nice example of thinking in terms of expressions. The scribe seems to have in mind a picture of the length being multiplied by three-fourths of the length, and the result being 12. Then the length squared has to be found by multiplying by what we would call the reciprocal of its coefficient, after which the length is found by taking the square root.

Slopes. The beginnings of trigonometry can be seen in Problems 56–60 of the papyrus, which involve the slope of the sides of pyramids and other figures. There is a unit of slope analogous to the *pesu* that we saw in Chapter 5 in the problems involving strength of bread and beer. The unit of slope is the *seked*, defined as the number of palms of horizontal displacement associated with a vertical displacement of 1 royal cubit. One royal cubit was 7 palms. Because of the relative sizes of

² Quoted by the second-century theologian Clement of Alexandria, in his *Miscellanies*, Book 1, Chapter 15.

horizontal and vertical displacements, it makes sense to use a larger unit of length for vertical distances than is used for horizontal distances, even at the expense of introducing an extra factor into computations of slope. In our terms the *seked* is seven times the tangent of the angle that the sloping side makes with the vertical. In some of the problems the *seked* is given in such a way that the factor of 7 drops out. Notice that if you were ordering a stone from the quarry, the *seked* would tell the stonecutter immediately where to cut. One would mark a point one cubit (distance from fingertip to elbow) from the corner in one direction and a point at a number of palms equal to the *seked* in the perpendicular direction, and then simply cut between the two points marked.

In Problem 57 a pyramid with a *seked* of $5 \frac{1}{4}$ and a base of 140 cubits is given. The problem is to find its height. The *seked* given here ($\frac{3}{4}$ of 7) is exactly that of one of the actual pyramids, the pyramid of Khafre, who reigned from 2558 to 2532 BCE. It appears that stones were mass-produced in several standard shapes with a *seked* that could be increased in intervals of one-fourth. Pyramid builders and designers could thereby refer to a standard brick shape, just as architects and contractors since the time of ancient Rome have been able to specify a standard diameter for a water pipe. Problem 58 gives the dimensions of the same pyramid and asks for its *seked*, apparently just to reinforce the reader's grasp of the relation between *seked* and dimension.

The circle. Five of the problems in the Ahmose Papyrus (41–43, 48, and 50) involve calculating the area of a circle. The answers given are approximations, but would be precise if the value $64/81$ used in the papyrus where we would use $\pi/4$ were exact. The author makes no distinction between the two. When physical objects such as grain silos are built, the parts used to build them have to be measured. In addition, the structures and their contents have a commercial, monetary value. Some number has to be used to express that value. It would therefore *not* be absurd—although it would probably be unnecessary—for a legislature to pass a bill prescribing a numerical value to be used for π .³ Similarly, the claim often made that the “biblical” value of π is 3, based on the description of a vat 10 cubits from brim to brim girdled by a line of 30 cubits (1 Kings 7:23) is pure pedantry. It assumes more precision than is necessary in the context. The author may have been giving measurements only to the nearest 10 cubits, not an unreasonable thing to do in a literary description.⁴

³ However, in the most notorious case where such a bill was nearly passed—House Bill 246 of the 1897 Indiana legislature—it *was* absurd. The bill was written by a physician and amateur mathematician named Edwin J. Goodwin. Goodwin had copyrighted what he thought was a quadrature of the circle. He offered to allow textbooks sold in Indiana to use his proof royalty-free provided that the Indiana House would pass this bill, whose text mostly glorified his own genius. Some of the mathematical statements the legislature was requested to enact were pure gibberish. For example, “a circular area is to the square on a line equal to the quadrant of the circumference, as the area of an equilateral rectangle is to the square on one side.” The one clear statement is that “the ratio of the chord and arc of ninety degrees... is as seven to eight.” That statement implies that $\pi = 16\sqrt{2}/7 \approx 3.232488$... The square root in this expression did not trouble Dr. Goodwin, who declared that $\sqrt{2} = 10/7$. At this point, one might have taken his value of π to be $160/49 = 3.265306122$... But, in a rare and uncalled-for manifestation of consistency, since he “knew” that $100/49 = (10/7)^2 = 2$, Goodwin declared this fraction equal to $16/5 = 3.2$. The bill was stopped at the last minute by lobbying from a member of the Indiana Academy of Sciences and was tabled without action.

⁴ However, like everything in the Bible, this passage has been subject to exhaustive and repeated analysis. For a summary of the conclusions reached in the Talmud, see Tsaban and Garber (1998).

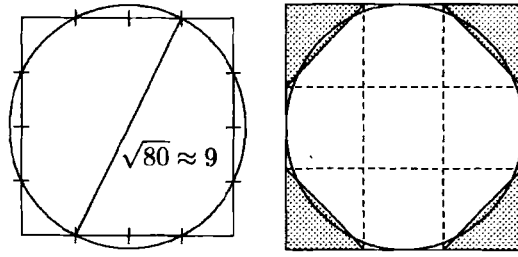


FIGURE 1. Conjectured explanations of the Egyptian squaring of the circle.

Ahmose takes the area of a circle to be the area of the square whose side is obtained by removing the ninth part of the diameter. In our language the area is the square on eight-ninths of the diameter, that is, it is the square on $\frac{16}{9}$ of the radius. In our language, not that of Egypt, this gives a value of π for area problems equal to $\frac{256}{81}$. Please remember, however, that the Egyptians had no concept of the number π . The constant of proportionality that they always worked with represents what we would call $\pi/4$. There have been various conjectures as to how the Egyptians might have arrived at this result. One such conjecture given by Robins and Shute (1987, p. 45) involves a square of side 8. If a circle is drawn through the points 2 units from each corner, it is visually clear that the four fillets at the corners, at which the square is outside the circle, are nearly the same size as the four segments of the circle outside the square; hence this circle and this square may be considered equal in area. Now the diameter of this circle can be obtained by connecting one of the points of intersection to the opposite point, as shown on the left-hand diagram in Fig. 1, and measurement will show that this line is very nearly 9 units in length (it is actually $\sqrt{80}$ in length). A second theory due to K. Vogel (see Gillings, 1972, pp. 143–144) is based on the fact that the circle inscribed in a square of side nine is roughly equal to the unshaded region in the right-hand diagram in Fig. 1. This area is $\frac{7}{9}$ of 81, that is, 63. A square of equal size would therefore have side $\sqrt{63} \approx 7.937 \approx 8$. In favor of Vogel's conjecture is the fact that a figure very similar to this diagram accompanies Problem 48 of the papyrus. A detailed discussion of various conjectures, giving connections with traditional African crafts, was given by Gerdes (1985).

The Pythagorean theorem. Inevitably in the discussion of ancient cultures, the question of the role played by the Pythagorean theorem is of interest. Did the ancient Egyptians know this theorem? It has been reported in numerous textbooks, popular articles, and educational videos that the Egyptians laid out right angles by stretching a rope with 12 equal intervals knotted on it so as to form a 3–4–5 right triangle. What is the evidence for this assertion? First, the Egyptians *did* lay out very accurate right angles. Also, as mentioned above, it is known that their surveyors used ropes as measuring instruments and were referred to as *rope-fixers* (see Plate 7). That is the evidence that was cited by the person who originally made the conjecture, the historian Moritz Cantor (1829–1920) in the first volume of his history of mathematics, published in 1882. The case can be made stronger, however. In his essay *Isis and Osiris* Plutarch says the following.

It has been imagined that the Egyptians regarded one triangle above all others, likening it to the nature of the universe. And in his *Republic* Plato seems to have used it in arranging marriages. This triangle has 3 on the vertical side, 4 on the base, and a hypotenuse of 5, equal in square to the other two sides. It is to be imagined then that it was constituted of the masculine on the vertical side, and the feminine on the base; also, Osiris as the progenitor, Isis as the receptacle, and Horus as the offspring. For 3 is the first odd number and is a perfect number; the 4 is a square formed from an even number of dyads; and the 5 is regarded as derived in one way from the father and another way from the mother, being made up of the triad and the dyad.

Still further, Berlin Papyrus 6619 contains a problem in which one square equals the sum of two others. It is hard to imagine anyone being interested in such conditions without knowing the Pythagorean theorem. Against the conjecture, we could note that the earliest Egyptian text that mentions a right triangle and finds the length of all its sides using the Pythagorean theorem dates from about 300 BCE, and by that time the presence of Greek mathematics in Alexandria was already established. None of the older papyri mention or use by implication the Pythagorean theorem.

On balance, one would guess that the Egyptians *did* know the Pythagorean theorem. However, there is no evidence that they used it to construct right angles, as Cantor conjectured. There are much simpler ways of doing that (even involving the stretching of ropes), which the Egyptians must have known. Given that the evidence for this conjecture is so meager, why is it so often reported as fact? Simply because it has been repeated frequently since it was originally made. We know precisely the source of the conjecture, but that knowledge does not seem to reach the many people who report it as fact.⁵

Spheres or cylinders? Problem 10 of the Moscow Papyrus has been subject to various interpretations. It asks for the area of a curved surface that is either half of a cylinder or a hemisphere. In either case it is worth noting that the area is obtained by multiplying the length of a semicircle by another length in order to obtain the area. Finding the area of a hemisphere is an extremely difficult problem. Intuitive techniques that work on flat or ruled surfaces break down, as shown in Problem 9.20. If the Egyptians did compute this area, no one has given any reasonable conjecture as to how they did so. The difficulty of this problem was given as one reason for interpreting the figure as half of a cylinder. Yet the plain language of the problem implies that the surface is a hemisphere. The problem was translated into German by the Russian scholar V. V. Struve (1889–1965); the following is a translation from the German:

The way of calculating a basket, if you are given a basket with an opening of $4\sqrt{2}$. O, tell me its surface!

⁵ This point was made very forcefully by van der Waerden (1963, p. 6). In his later book, *Geometry and Algebra in Ancient Civilizations*, van der Waerden claimed that integer-sided right triangles, which seem to imply knowledge of the Pythagorean theorem, are ubiquitous in the oldest megalithic structures. Thus, he seems to imply that the Egyptians knew the theorem, but didn't use it as Cantor suggested.

Calculate $\bar{9}$ of 9, since the basket is half of an egg. The result is 1. Calculate what is left as 8. Calculate $\bar{9}$ of 8. The result is $\bar{3} \bar{6}$ $\bar{18}$. Calculate what is left of this 8 after this $\bar{3} \bar{6} \bar{18}$ is taken away. The result is 7 $\bar{9}$. Calculate 4 $\bar{2}$ times with 7 $\bar{9}$. The result is 32. Behold, this is the surface. You have found it correctly.

If we interpret the basket as being a hemisphere, the scribe has first doubled the diameter of the opening from 4 $\bar{2}$ to 9 “because the basket is half of an egg.” (If it had been the *whole* egg, the diameter would have been quadrupled.) The procedure used for finding the area here is equivalent to the formula $2d \cdot \frac{8}{9} \cdot \frac{8}{9} \cdot d$. Taking $(\frac{8}{9})^2$ as representing $\pi/4$, we find it equal to $(\pi d^2)/2$, or $2\pi r^2$, which is indeed the area of a hemisphere of radius r .

This value is also the lateral area of half of a cylinder of height d and base diameter d . If the basket is interpreted as half of a cylinder, the opening would be square and the number 4 $\bar{2}$ would be the side of the square. That would mean also that the “Egyptian π ” ($\pi/4 = 64/81$), used for area problems was also being applied to the ratio of the circumference to the diameter. The numerical answer is consistent with this interpretation, but it does seem strange that only the lateral surface of the cylinder was given. That would indicate that the basket was open at the sides. It would be strange to describe such a basket as “half of an egg.” The main reason given by van der Waerden (1963, pp. 33–34) for preferring this interpretation is an apparent inaccuracy in Struve’s statement of the problem. Van der Waerden quotes T. E. Peet, who says that the number 4 $\bar{2}$ occurs twice in the statement of the problem, as the opening of the top of the basket and also as its depth. This interpretation, however, leads to further difficulties. If the surface is indeed half of a cylinder of base diameter 4 $\bar{2}$, its depth is not 4 $\bar{2}$; it is 2 $\bar{4}$. Van der Waerden also mentions a conjecture of Neugebauer, that this surface was intended to be a domelike structure of a sort seen in some Egyptian paintings, resembling very much the small end of an egg. That interpretation restores the idea that this problem was the computation of the area of a nonruled surface, and the approximation just happens to be the area of a hemisphere.

1.2. Volumes. One of the most remarkable achievements of the Egyptians is the discovery of accurate ways of computing volumes. As in the case of surface areas, the most remarkable result is found in the Moscow, not the Ahmose, Papyrus. In Problem 41 of the Ahmose Papyrus we find the correct procedure used for finding the volume of a cylindrical silo, that is, the area of the circular base is multiplied by the height. To make the numbers easy, the diameter of the base is given as 9 cubits, as in Problems 48 and 50, so that the area is 64 square cubits. The height is 10 cubits, giving a volume of 640 cubic cubits. However, the standard unit of grain volume was a *khar*, which is two-thirds of a cubic cubit, resulting in a volume of 960 *khar*. In a further twist, to get a smaller answer, the scribe divides this number by 20, getting 48 “hundreds of quadruple *hekats*.” (A *khar* was 20 *hekats*.) Problem 42 is the same problem, only with a base of diameter 10 cubits. Apparently, once the reader has the rule well in hand, it is time to test the limits by making the data more cumbersome. The answer is computed to be 1185 $\bar{6}$ $\bar{54}$ *khar*, again expressed in hundreds of quadruple *hekats*. Problems 44–46 calculate the volume of prisms on a rectangular base by the same procedure.

Given that pyramids are so common in Egypt, it is surprising that the Ahmose Papyrus does not discuss the volume of a pyramid. However, Problem 14 from the Moscow Papyrus asks for the volume of the frustum of a square pyramid given that the side of the lower base is 4, the side of the upper base is 2, and the height is 6. The author gives the correct recipe: Add the areas of the two bases to the area of the rectangle whose sides are the sides of the bases, that is, $2 \cdot 2 + 4 \cdot 4 + 2 \cdot 4$, then multiply by one-third of the height, getting the correct answer, 56. This technique could not have been arrived at through experience. Some geometric principle must be involved, since the writer knew that the sides of the bases, which are *parallel* lines, need to be multiplied. Normally, the lengths of two lines are multiplied only when they are perpendicular to each other, so that the product represents the area of a rectangle. Gillings (1972, pp. 190–193) suggests a possible route. Robins and Shute (1987, pp. 48–49) suggest that the result may have been obtained by completing the frustum to a full pyramid, and then subtracting the volume of the smaller pyramid from the larger. In either case, the power of visualization involved in seeing that the relation is the correct one is remarkable.

Like the surface area problem from the Moscow Papyrus just discussed, this problem reflects a level of geometric insight that must have required some accumulation of observations built up over time. It is very easy to see that if a right pyramid with a square base is sliced in half by a plane through its vertex and a pair of diagonally opposite vertices of the base, the base is bisected along with the pyramid. Thus, a tetrahedron whose base is half of a square has volume exactly half that of the pyramid of the same height having the whole square as a base.

It is also easy to visualize how a cube can be cut into two wedges, as in the top row of Fig. 2. Each of these wedges can then be cut into a pyramid on a face of the cube plus an extra tetrahedron, as in the bottom row. The tetrahedron $P'Q'R'S'$ has a base that is half of the square base of the pyramid $PQRST$, and hence has half of its volume. It follows that the volume of the tetrahedron is one-sixth that of the cube, and so the pyramid $PQRST$ is one-third of the volume. A “mixed” strategy is also possible, involving weighing of the parts. The two tetrahedra would, in theory, balance one of the square pyramids. This model could be sawn out of stone or wood. From that special case one might generalize the vital clue that the volume of a pyramid is one-third the area of the base times the altitude.

Once the principle is established that a pyramid equals a prism on the same base with one-third the height, it is not difficult to chop a frustum of a pyramid into the three pieces described in the Moscow Papyrus. Referring to Fig. 3, which shows a frustum with bottom base a square of side a and upper base a square of side b with $b < a$, we can cut off the four corners and replace them by four rectangular solids with square base of side $(a - b)/2$ and height $h/3$. These four fit together to make a single solid with square base of side $a - b$ and height $h/3$. One opposite pair of the four sloping faces that remain after the corners are removed can be cut off, turned upside down, and laid against the other two sloping faces so as to make a single slab with a rectangular base that is $a \times b$ and has height h . The top one-third of this slab can then be cut off and laid aside. It has volume $(h/3)ab$. The top half of what remains can then be cut off, and a square prism of side b and height $h/3$ cut off from it. If that square prism is laid aside (it has volume $(h/3)b^2$), the remaining piece, which is $(a - b) \times b \times (h/3)$, will fill out the other corner of the bottom layer, resulting in a square prism of volume $(h/3)a^2$. Thus, we obtain the

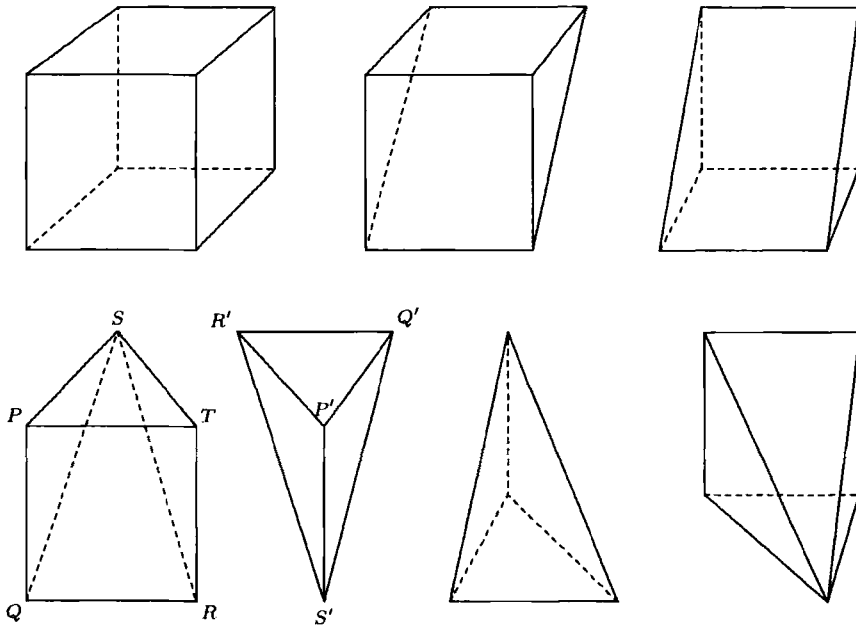


FIGURE 2. Dissection of a cube into two square pyramids and two tetrahedra.

three pieces that the scribe added to get the volume of the frustum in a way that is not terribly implausible.

It goes without saying that the last few paragraphs and Figs. 2 and 3 are conjectures, not facts of history. We do not know how the Egyptians discovered that the volume of a pyramid is one-third the volume of a prism of the same base and height or how they found the volume of a frustum. The little story just presented is merely one possible scenario.

2. Mesopotamia

Mesopotamian geometry, like its Egyptian counterpart, was regarded more as an application of mathematics than as mathematics proper. The primary emphasis was on areas and volumes. However, the Mesopotamian tablets suggest a very strong algebraic component. Many of the problems that are posed in geometric garb have no apparent practical application but are very good exercises in algebra. For example, British Museum tablet 13901 contains the following problem: *Given two squares such that the side of one is two-thirds that of the other plus 5 GAR and whose total area is 25,25 square GAR, what are the sides of the squares?* Where in real life would one encounter such a problem? The tablet itself gives no practical context, and we conclude that this apparently geometric problem is really a problem in algebraic manipulation of expressions. As Neugebauer states (1952, p. 41), "It is easy to show that geometrical concepts play a very secondary part in Babylonian algebra, however extensively a geometrical terminology may be used." Both Neugebauer and van der Waerden (1963, p. 72) point out that the cuneiform tablets contain operations that are geometrically absurd, such as adding

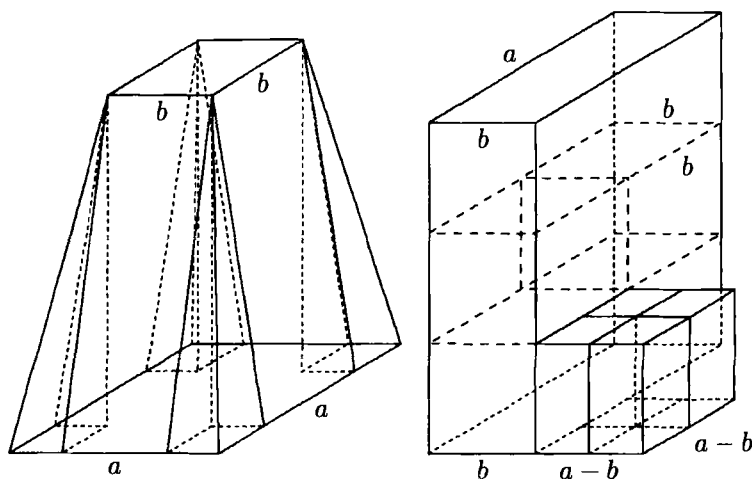


FIGURE 3. Dissection of a frustum of a pyramid.

a length to an area or multiplying two areas. For that reason, our discussion of such problems is postponed to Chapter 13.

2.1. The Pythagorean theorem. In contrast to the case of Egypt, there is clear proof that the Mesopotamians knew the Pythagorean theorem in full generality at least 1000 years before Pythagoras. They were thus already on the road to finding more abstract properties of geometric figures than mere size. Of course, this theorem was known at an early date in India and China, so that one cannot say certainly where the earliest discovery was and whether the appearance of this theorem in different localities was the result of independent discovery or transmission. But as far as present knowledge goes, the earliest examples of the use of the “Pythagorean” principle that the square on the hypotenuse of a right triangle equals the sum of the squares on the other two legs occur in the cuneiform tablets. Specifically, the old Babylonian text known as BM 85 196 contains a problem that has appeared in algebra books for centuries. We give it below as Problem 9.4. In this problem we are dealing with a right triangle of hypotenuse 30 with one leg equal to $30 - 6 = 24$. Obviously, this is the famous 3-4-5 right triangle with all sides multiplied by 6. Obviously also, the interest in this theorem was more numerical than geometric. How often, after all, are we called upon to solve problems of this type in everyday life?

How might the Pythagorean theorem have been discovered? The following hypothesis was presented by Allman (1889, pp. 35-37), who cited a work (1870) by Carl Anton Bretschneider (1808-1878). Allman thought this dissection was due to the Egyptians, since, he said, it was done in their style. If he was right, the Egyptians did indeed discover the theorem.

Suppose that you find it necessary to construct a square twice as large as a given square. How would you go about doing so? (This is a problem the Platonic Socrates poses in the dialogue *Meno*.) You might double the side of the square, but you would soon realize that doing so actually quadruples the size of the square. If you drew out the quadrupled square and contemplated it for a while, you might be

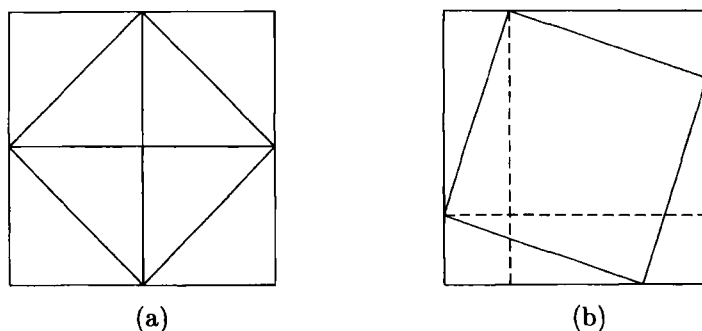


FIGURE 4. (a) Doubling a square; (b) the Pythagorean theorem.

led to join the midpoints of its sides in order, that is, to draw the diagonals of the four copies of the original square. Since these diagonals cut the four squares in half, they will enclose a square twice as big as the original one (Fig. 4). It is quite likely that someone, either for practical purposes or just for fun, discovered this way of doubling a square. If so, someone playing with the figure might have considered the result of joining in order the points at a given distance from the corners of a square instead of joining the midpoints of the sides. Doing so creates a square in the center of the larger square surrounded by four copies of a right triangle whose hypotenuse equals the side of the center square (Fig. 4); it also creates the two squares on the legs of that right triangle and two rectangles that together are equal in area to four copies of the triangle. (In Fig. 4 one of these rectangles is divided into two equal parts by its diagonal, which is the hypotenuse of the right triangle.) Hence the larger square consists of four copies of the right triangle plus the center square. It also consists of four copies of the right triangle plus the squares on the two legs of the right triangle. The inevitable conclusion is that *the square on the hypotenuse of any right triangle equals the sum of the squares on the legs*. This is the Pythagorean theorem, and it is used in many places in the cuneiform texts.

2.2. Plane figures. Some cuneiform tablets give the area of a circle of unit radius, which we have called the two-dimensional π , as 3. On the other hand (Neugebauer, 1952, p. 46), the one-dimensional π was known to slightly more accuracy. On a tablet excavated at Susa in 1936, it was stated that the perimeter of a regular hexagon, which is three times its diameter, is 0;57,36 times the circumference of the circumscribed circle. That makes the circumference of a circle of unit diameter equal to

$$\frac{3}{0;57,36} = \frac{25}{8} = 3.125.$$

That the Mesopotamian mathematicians recognized the relation between the area and the circumference of a circle is shown by two tablets from the Yale Babylonian Collection (YBC 7302 and YBC 11120, see Robson, 2001, p. 180). The first contains a circle with the numbers 3 and 9 on the outside and 45 on the inside. These numbers fit perfectly the formula $A = C^2/(4\pi)$, given that the scribe was using $\pi = 3$. Assuming that the 3 represents the circumference, 9 its square and 45 the quotient, we find $9/(4 \cdot 3) = 3/4 = 0;45$. Confirmation of this hypothesis comes from the other tablet, which contains 1;30 outside and 11;15 inside, since $(1;30^2)/(4 \cdot 3) = (2;15)/12 = 135/12 = 11.25 = 11;15$.

The strongest area of Mesopotamian science that has been preserved is astronomy, and it is here that geometry becomes most useful. The measurement of angles—arcs of circles—is essential to observation of the Sun, Moon, stars, and planets, since to the human eye they all appear to be attached to a large sphere rotating overhead. The division of the circle into 360 degrees is one convention that came from Mesopotamia, was embraced by the Greeks, and became an essential part of applied geometry down to the present day. The reason for the number 360 is the base-60 computational system used in Mesopotamia. The astronomers divided all circles into 360 or 720 equal parts and the radius into 60 equal parts. In that way, a unit of length along the radius was approximately equal to a unit of length on the circle.

2.3. Volumes. The cuneiform tablets contain computations of some of the same volumes as the Egyptian papyri. For example, the volume of a frustum of a square pyramid is computed in an old Babylonian tablet (British Museum 85 194). This volume is computed correctly in the Moscow Papyrus, but the Mesopotamian scribe seems to have generalized incorrectly from the case of a trapezoid and reasoned that the volume is the height times the average area of the upper and lower faces. This rule overestimates the volume by twice the volume of the four corners cut out in Fig. 3. There is, however, some disagreement as to the correct translation of the tablet in question. Neugebauer (1935, Vol. 1, p. 187) claimed that the computation was based on an algebraic formula that is geometrically correct. The square bases are given as having sides 10 and 7 respectively, and the height is given as 18. The incorrect rule we are assuming would give a volume of 1341, which is 22,21 in sexagesimal notation; but the actual text reads 22,30. The discrepancy could be a simple misprint, with three ten-symbols carelessly written for two ten-symbols and a one-symbol. The computation used is not entirely clear. The scribe first took the average base side $(10 + 7)/2$ and squared it to get 1,12;15 in sexagesimal notation (72.25). At this point there is apparently some obscurity in the tablet itself. Neugebauer interpreted the next number as 0;45, which he assumed was calculated as one-third of the square of $(10 - 7)/2$. The sum of these two numbers is 1,13, which, multiplied by 18, yields 21,54 (that is, 1314), which is the correct result. But it is difficult to see how this number could have been recorded incorrectly as 22,30. If the number that Neugebauer interprets as 0;45 is actually 2;15 (which is a stretch—three ten-symbols would have to become two one-symbols), it would be exactly the square of $(10 - 7)/2$, and it would yield the same incorrect formula as the assumption that the average of the areas of the two bases was being taken. In any case, the same procedure is used to compute the volume of the frustum of a cone (Neugebauer, 1935, p. 176), and in that case it definitely is the incorrect rule stated here, taking the average of the two bases and multiplying by the height.

3. China

Three early Chinese documents contain a considerable amount of geometry, always connected with the computation of areas and volumes. We shall discuss the geometry in them in chronological order, omitting the parts that repeat procedures we have already discussed in connection with Egyptian geometry.

3.1. The *Zhou Bi Suan Jing*. As mentioned in Chapter 2, the earliest Chinese mathematical document still in existence, the *Zhou Bi Suan Jing*, is concerned

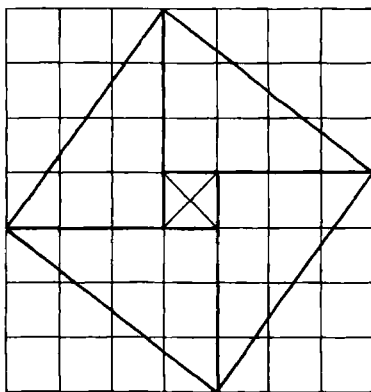


FIGURE 5. Chinese illustration of the Pythagorean theorem.

with astronomy and the applications of mathematics to the study of the heavens. The title refers to the use of the sundial or gnomon in astronomy. This is the physical model that led the Chinese to discover the Pythagorean theorem. Here is a paraphrase of the discussion:

Cut a rectangle whose width is 3 units and whose length is 4 (units) along its diagonal. After drawing a square on this diagonal, cover it with half-rectangles identical to the piece of the original rectangle that lies outside the square, so as to form a square of side 7. [See Fig. 5.] Then the four outer half-rectangles, each of width 3 and length 4 equal two of the original rectangle, and hence have area 24. When this amount is subtracted from the square of area 49, the difference, which is the area of the square on the diagonal, is seen to be 25. The length of the diagonal is therefore 25.

Although the proof is given only for the easily computable case of the 3-4-5 right triangle, it is obvious that the geometric method is perfectly general, lacking only abstract symbols for unspecified numbers. In our terms, the author has proved that the length of the diagonal of a rectangle whose width is a and whose length is b is the square root of $(a + b)^2 - 2ab$. Note that this form of the theorem is not the " $a^2 + b^2 = c^2$ " that we are familiar with.

The *Zhou Bi Suan Jing* contains three diagrams accompanying the discussion of the Pythagorean theorem. According to Cullen (1996, p. 69), one of these diagrams was apparently added in the third century by the commentator Zhao Shuang. This diagram is shown in Fig. 5 for the special case of a 3-4-5 triangle. The other two were probably added by later commentators in an attempt to elucidate Zhao Shuang's commentary.

According to Li and Du (1987, p. 29), the vertical bar on a sundial was called *gu* in Chinese, and its shadow on the sundial was called *gou*; for that reason the Pythagorean theorem was known as the *gougu* theorem. Cullen (1996, p. 77) says that *gu* means *thigh* and *gou* means *hook*. All authorities agree that the hypotenuse was called *xian* (bowstring). The *Zhou Bi Suan Jing* says that the Emperor Yu was able to bring order into the realm because he knew how to use this theorem to compute distances. Zhao Shuang credited the Emperor Yu with saving his people from floods and other great calamities, saying that in order to do so he had to

survey the shapes of mountains and rivers. Apparently the Emperor had drainage canals dug to channel floods out of the valleys and into the Yangtze and Yellow Rivers.

The third-century commentary on the *Zhou Bi Suan Jing* by Zhao Shuang explains a method of surveying that was common in China, India, and the Muslim world for centuries. The method is illustrated in Fig. 6, which assumes that the height H of an inaccessible object is to be determined. To determine H , it is necessary to put two poles of a known height h vertically into the ground in line with the object at a known distance D apart. The height h and the distance D are theoretically arbitrary, but the larger they are, the more accurate the results will be. After the poles are set up, the lengths of the shadows they would cast if the Sun were at the inaccessible object are measured as s_1 and s_2 . Thus the lengths s_1 , s_2 , h , and D are all known. A little trigonometry and algebra will show that

$$H = h + \frac{Dh}{s_2 - s_1}.$$

We have given the result as a formula, but as a set of instructions it is very easy to state in words: *The required height is found by multiplying the height of the poles by the distance between them, dividing by the difference of the shadow lengths, and adding the height of the poles.*

This method was expounded in more detail in a commentary on the *Jiu Zhang Suanshu* written by Liu Hui in 263 CE. This commentary, along with the rest of the material on right triangles in the *Jiu Zhang Suanshu* eventually became a separate treatise, the *Hai Dao Suan Jing* (*Sea Island Mathematical Manual*, see Ang and Swetz, 1986). Liu Hui mentioned that this method of surveying could be found in the *Zhou Bi Suan Jing* and called it the *double difference method* (*chong cha*). The name apparently arises because the difference $H - h$ is obtained by dividing Dh by the difference $s_2 - s_1$.

We have described the lengths s_1 and s_2 as shadow lengths here because that is the problem used by Zhao Shuang to illustrate the method of surveying. He attempts to calculate the height of the Sun, given that at the summer solstice a stake 8 *chi* high casts a shadow 6 *chi* long and that the shadow length decreases by 1 *fen* for every 100 *li* that the stake is moved south, casting no shadow at all when moved 60,000 *li* to the south. This model assumes a flat Earth, under which the shadow length is proportional to the distance from the pole to the foot of the perpendicular from the Sun to the plane of the Earth. Even granting this assumption, as we know, the Sun is so distant from the Earth that no lengthening or shortening of shadows would be observed. To any observable precision the Sun's rays are parallel at all points on the Earth's surface. The small change in shadow length that we observe is due entirely to the curvature of the Earth. But let us continue, accepting Zhao Shuang's assumptions.

The data here are $D = 1000$ *li*, $s_2 - s_1 = 1$ *fen*, $h = 8$ *chi*. One *chi* is about 25 centimeters, one *fen* is about 2.5 cm, and one *li* is 1800 *chi*, that is, about 450 meters. Because the pole height h is obviously insignificant in comparison with the height of the Sun, we can neglect the first term in the formula we gave above, and write

$$H = \frac{Dh}{s_2 - s_1}.$$

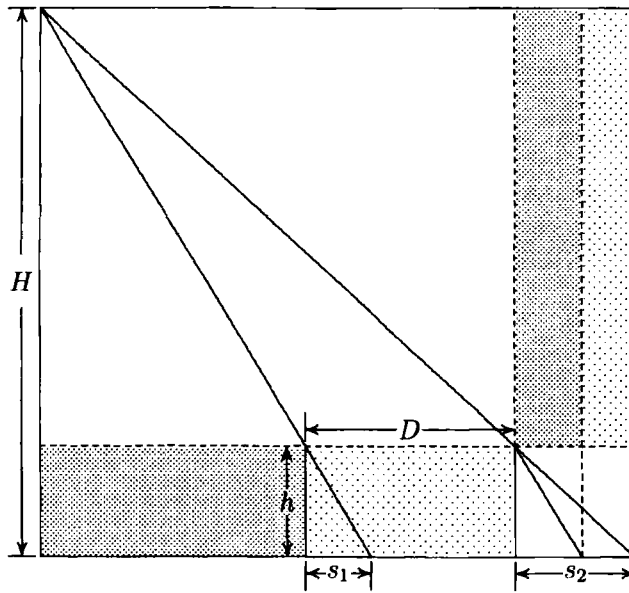


FIGURE 6. The double-difference method of surveying.

When we insert the appropriate values, we find, as did Zhao Shuang, that the Sun is 80,000 *li* high, about 36,000 kilometers. Later Chinese commentators recognized that this figure was inaccurate, and in the eighth century an expedition to survey accurately a north-south line found the actual lengthening of the shadow to be 4 *fen* per thousand *li*. Notice that there seem to be two methods of computing the height here. In the method just discussed, the fact that the Sun is directly overhead at a distance of 60,000 *li* to the south is irrelevant to the computation. If it is taken into account, one can immediately use the similar triangles to infer the height of 80,000 *li*. This fact suggests that the original text was modified by later commentators, but that not all the parts that became irrelevant as a result of the modifications were removed.

3.2. The *Jiu Zhang Suanshu*. The *Jiu Zhang Suanshu* contains all the standard formulas for the areas of squares, rectangles, triangles, and trapezoids, and also the recognition of a relation between the circumference and the area of a circle, which we could interpret as a connection between the one-dimensional π and the two-dimensional π . The geometric formulas given in this treatise are more extensive than those of the Ahmose Papyrus; for example, there are approximate formulas for the volume of segment of a sphere and the area of a segment of a circle. It is perhaps not fair to compare the two documents, since the Ahmose Papyrus was written nearly two millennia earlier, and the *Jiu Zhang Suan Shu* was intended to cover all the mathematics known at the time. The implied value of one-dimensional π , however, is $\pi = 3$. It is surprising to find this value so late, since it is known that the value 3.15147 had been obtained in China by the first century. According to Li and Du (1987, p. 68), Liu Hui refined it to $3.14 + 64/62500 = 3.141024$ by

approximating the area of a 192-sided polygon.⁶ That is, he started with a hexagon and doubled the number of sides five times. However, since the area of the polygon with twice the number of sides is the radius of the circumscribed circle times the perimeter of the original polygon, it was only necessary to find the perimeter of a 96-sided polygon and multiply by the radius.

Problems 31 and 32 ask for the area of a circular field of a given diameter and circumference.⁷ The method is to multiply half of the circumference by half of the diameter, which is exactly right in terms of Euclidean geometry; equivalently, the reader is told that one may multiply the two quantities and divide by 4. However, in the actual data for problems the diameter given is exactly one-third of the given circumference; in other words, the value assumed for one-dimensional π is 3. The assumption of that value leads to two other procedures for calculating the area: squaring the diameter, then multiplying by 3 and dividing by 4, or squaring the circumference and dividing by 12. An elaboration of this problem occurs in Problems 37 and 38, in which the area of an annulus (the region outside the smaller of two concentric circles and inside the larger) is given in terms of its width and the circumferences of the two circles.

The authors knew also how to find the volume of a pyramid. Problem 15 of Chapter 5 asks for the volume of a pyramid whose base is a rectangle 5 *chi* by 7 *chi* and whose height is 8 *chi*. The answer is given as $93\frac{1}{3}$ (cubic) *chi*. For a frustum of a pyramid having rectangular bases the recipe is to add twice the length of the upper base to the lower base and multiply by the width of the upper base to get one term. A second term is obtained symmetrically as twice the length of the lower base plus the length of the upper base, multiplied by the width of the lower base. These two terms are then added and multiplied by the height, after which one divides by 6. If the bases are $a \times b$ and $c \times d$ (the sides of length a and c being parallel) and the height is h , this yields what we would write (correctly) as

$$V = \frac{h}{6} [(2a + c)b + (2c + a)d].$$

Notice that this result is more general than the formula in the Moscow Papyrus, which is given for a frustum with square bases.

The Pythagorean theorem. The last of the nine chapters of the *Jiu Zhang Suanshu* contains 24 problems on the *gougu* theorem. After a few “warm-up” problems in which two of the three sides of a right triangle are given and the third is to be computed, the problems become more complicated. Problem 11, for example, gives a rectangular door whose height exceeds its width by 6 *chi*, 8 *cun* and has a diagonal of 1 *zhang*. One *zhang* is 10 *chi* and 1 *chi* is 10 *cun* (apparently a variant rendering of *fen*). The recipe given is correct: Take half the difference of the height and width, square it, double, subtract from the square of the diagonal, then take the square root of half of the result. That process yields the average of the height and width, and given their semidifference of 3 *chi*, 4 *cun*, one can easily get both the width and the height.

3.3. The *Sun Zi Suan Jing*. The *Sun Zi Suan Jing* contains a few problems in measurement that are unusual enough to merit some discussion. An inverse area

⁶ Lam and Ang (1986) give the value as $3.14 + 169/625 = 3.142704$.

⁷ All references to problem numbers and nomenclature in this section are based on the article of Lam (1994).



FIGURE 7. The double square umbrella.

problem occurs in Problem 20, in which a circle is said to have area 35,000 square *bu*, and its circumference is required. Since the area is taken as one-twelfth of the square of the circumference, the author multiplies by 12, then takes the square root, getting $648\frac{96}{1296}$ *bu*.

3.4. Liu Hui. Chinese mathematics was greatly enriched from the third through the sixth centuries by a series of brilliant geometers, whose achievements deserve to be remembered alongside those of Euclid, Archimedes, and Apollonius. We have space to discuss only three of these, beginning with the third-century mathematician Liu Hui (ca. 220–ca. 280). Liu Hui had a remarkable ability to visualize figures in three dimensions. In his commentary on the *Jiu Zhang Suanshu* he asserted that the circumference of a circle of diameter 100 is 314. In solid geometry he provided dissections of many geometric figures into pieces that could be reassembled to demonstrate their relative sizes beyond any doubt. As a result, real confidence could be placed in the measurement formulas that he provided. He gave correct procedures, based on such dissections, for finding the volumes enclosed by many different kinds of polyhedra. But his greatest achievement is his work on the volume of the sphere.

The *Jiu Zhang Suanshu* made what appears to be a very reasonable claim: that the ratio of the volume enclosed by a sphere to the volume enclosed by the circumscribed cylinder can be obtained by slicing the sphere and cylinder along the axis of the cylinder and taking the ratio of the area enclosed by the circular cross section of the sphere to the area enclosed by the square cross section of the cylinder. In other words, it would seem that the ratio is $\pi : 4$. This conjecture seems plausible, since every such section produces exactly the same figure. It fails, however because of what is called *Pappus' principle*: The volume of a solid of revolution equals the area revolved about the axis times the distance traveled by the centroid of the area. The half of the square that is being revolved to generate the cylinder has a centroid that is farther away from the axis than the centroid of the semicircle inside it whose revolution produces the sphere; hence when the two areas are multiplied by the two distances, their ratios get changed. When a circle inscribed in a square

is rotated, the ratio of the volumes generated is 2:3, while that of the original areas is $\pi : 4$. Liu Hui noticed that the sections of the figure parallel to the base of the cylinder do not all have the same ratios. The sections of the cylinder are all disks of the same size, while the sections of the sphere shrink as the section moves from the equator to the poles. He also formed a solid by intersecting two cylinders circumscribed about the sphere whose axes are at right angles to each other, thus producing a figure he called a *double square umbrella*, which is now known as a *bicylinder* or *Steinmetz solid* (see Hogendijk, 2002). A representation of the double square umbrella, generated using *Mathematica* graphics, is shown in Fig. 7. Its volume *does* have the same ratio to the sphere that the square has to its inscribed circle, that is, $4 : \pi$. This proportionality between the double square umbrella and the sphere is easy to see intuitively, since every horizontal slice of this figure by a plane parallel to the plane of the axes of the two circumscribed cylinders intersects the double square umbrella in a square and intersects the sphere in the circle inscribed in that square. Liu Hui inferred that the volume enclosed by the double umbrella would have this ratio to the volume enclosed by the sphere. This inference is correct and is an example of what is called *Cavalieri's principle*: *Two solids such that the section of one by each horizontal plane bears a fixed ratio to the section of the other by the same plane have volumes in that same ratio*. This principle had been used by Archimedes five centuries earlier, and in the introduction to his *Method*, Archimedes uses *this very example*, and asserts that the volume of the intersection of the two cylinders is two-thirds of the volume of the cube in which they are inscribed.⁸ But Liu Hui's use of it (see Lam and Shen, 1985) was obviously independent of Archimedes. It amounts to a limiting case of the dissections he did so well. The solid is cut into *infinitely thin* slices, each of which is then dissected and reassembled as the corresponding section of a different solid. This realization was a major step toward an accurate measurement of the volume of a sphere. Unfortunately, it was not granted to Liu Hui to complete the journey. He maintained a consistent agnosticism on the problem of computing the volume of a sphere, saying, "Not daring to guess, I wait for a capable man to solve it."

3.5. Zu Chongzhi. That "capable man" required a few centuries to appear, and he turned out to be two men. "He" was Zu Chongzhi (429–500) and his son Zu Geng (450–520). Zu Chongzhi was a very capable geometer and astronomer who said that if the diameter of a circle is 1, then the circumference lies between 3.1415926 and 3.1415927. From these bounds, probably using the Chinese version of the Euclidean algorithm, the method of mutual subtraction (see Problem 7.12), he stated that the circumference of a circle of diameter 7 is (approximately) 22 and that of a circle of diameter 113 is (approximately) 355.⁹ These estimates are very good, far too good to be the result of any inspired or hopeful guess. Of course, we don't have to imagine that Zu Chongzhi actually *drew* the polygons. It suffices to know how to compute the perimeter, and that is a simple recursive process: If s_n is the length of the side of a polygon of n sides inscribed in a circle of unit radius, then

$$s_{2n}^2 = 2 - \sqrt{4 - s_n^2}.$$

⁸ Hogendijk (2002) argues that Archimedes also knew the surface area of the bicylinder.

⁹ The approximation $\pi \approx \frac{22}{7}$ was given earlier by He Chengtian (370–447), and of course much earlier by Archimedes. A more sophisticated approach by Zhao Youqin (b. 1271) that gives $\frac{355}{113}$ was discussed by Volkov (1997).

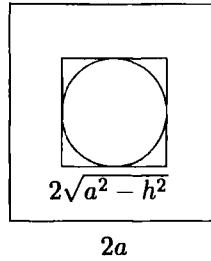


FIGURE 8. Sections of the cube, double square umbrella, and sphere at height h .

Hence each doubling of the number of sides makes it necessary to compute a square root, and the approximation of these square roots must be carried out to many decimal places in order to get enough guard digits to keep the errors from accumulating when you multiply this length by the number of sides. In principle, however, given enough patience, one could compute any number of digits of π this way.

One of Zu Chongzhi's outstanding achievements, in collaboration with his son, was finding the volume enclosed by Liu Hui's double square umbrella. As Fu (1991) points out, this volume was not approachable by the direct method of dissection and recombination that Liu Hui had used so successfully.¹⁰ An indirect approach was needed. The trick turned out to be to enclose the double square umbrella in a cube and look at the volume inside the cube and outside the double square umbrella. Suppose that the sphere has radius a . The double square umbrella can then be enclosed in a cube of side $2a$. Consider a horizontal section of the enclosing cube at height h above the middle plane of that cube. In the double umbrella this section is a square of side $2\sqrt{a^2 - h^2}$ and area $4(a^2 - h^2)$, as shown in Fig. 8. Therefore the area outside the double umbrella and inside the cube is $4h^2$.

It was no small achievement to look at the region in question. It was an even keener insight on the part of the family Zu to realize that this cross-sectional area is equal to the area of the cross section of an upside-down pyramid with a square base of side $2a$ and height a . Hence *the volume of the portion of the cube outside the double umbrella in the upper half of the cube equals the volume of a pyramid with square base $2a$ and height a* . But thanks to earlier work contained in Liu Hui's commentaries on the *Jiu Zhang Suanshu*, Zu Chongzhi knew that this volume was $(4a^3)/3$. It therefore follows, after doubling to include the portion below the middle plane, that the region inside the cube but outside the double umbrella has volume $(8a^3)/3$, and hence that the double umbrella itself has volume $8a^3 - (8a^3)/3 = (16a^3)/3$.

Since, as Liu Hui had shown, the volume of the sphere is $\pi/4$ times the volume of the double square umbrella, it follows that the sphere has volume $(\pi/4) \cdot (16a^3)/3$, or $(4\pi a^3)/3$.

¹⁰ Lam and Shen (1985, p. 223), however, say that Liu Hui *did* consider the idea of setting the double umbrella inside the cube and trying to find the volume between the two. Of course, that volume also is not accessible through direct, finite dissection.

4. Japan

The *Wasanists* mentioned in Chapter 3, whose work extended from 1600 to 1850, inherited a foundation of mathematics established by the great Chinese mathematicians, such as Liu Hui, Zu Chongzhi, and Yang Hui. They had no need to work out procedures for computing the areas and volumes of simple figures. The only problems in elementary measurement of figures that had not been solved were those involving circles and spheres, connected, as we know, with the value of π in various dimensions. Nevertheless, during this time there was a strong tradition of geometric challenge problems. It has already been mentioned that religious shrines in Japan were frequently decorated with the solutions of such problems (see Plate 2). The geometric problems that were solved usually involved combinations of simple figures whose areas or volumes were known but which were arranged in such a way that finding their parts became an intricate problem in algebra. The word *algebra* needs to be emphasized here. The challenge in these problems was only incidentally geometric; it was largely algebraic, as the book of Fukagawa and Pedoe (1989) shows very convincingly. New geometry arose in Japan near the end of the seventeenth century, with better approximations to π and the solution of problems involving the rectification of arcs and the computation of the volume and area of a sphere by methods using infinite series and sums that approximate integrals.

We begin by mentioning a few of the challenge problems without giving their solutions, since they are really problems in algebra. Afterward we shall briefly discuss the infinitesimal methods used to solve the problems of measuring arcs, areas, and volumes in spheres.

4.1. The challenge problems. In 1627 Yoshida Koyu wrote the *Jinkō-ki* (*Treatise on Large and Small Numbers*), concluding it with a list of challenge questions, and thereby stimulated a great deal of further work. Here are some of the questions:

1. There is a log of precious wood 18 feet long whose bases are 5 feet and $2\frac{1}{2}$ feet in circumference. Into what lengths should it be cut to trisect the volume?
2. There have been excavated 560 measures of earth, which are to be used for the base of a building. The base is to be 3 measures square and 9 measures high. Required, the size of the upper base.
3. There is a mound of earth in the shape of a frustum of a circular cone. The circumferences of the bases are 40 measures and 120 measures and the mound is 6 measures high. If 1200 measures of earth are taken evenly off the top, what will be the height?
4. A circular piece of land 100 [linear] measures in diameter is to be divided among three persons so that they shall receive 2900, 2500, and 2500 [square] measures respectively. Required, the lengths of the chords and the altitudes of the segments.

These problems were solved in a later treatise, which in turn posed new mathematical problems to be solved; this was the beginning of a tradition of posing and solving problems that lasted for 150 years. Seki Kōwa solved a geometric problem that would challenge even the best algebraist today. It was the fourteenth in a list of challenge problems posed by Sawaguchi Kazuyuki: *There is a quadrilateral whose sides and diagonals are u , v , w , x , y , and z* [as shown in Fig. 9].

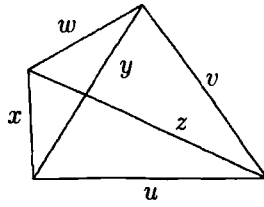


FIGURE 9. Sawaguchi Kazuyuki's quadrilateral problem.

It is given that

$$\begin{aligned} z^3 - u^3 &= 271 \\ u^3 - v^3 &= 217 \\ v^3 - y^3 &= 60.8 \\ y^3 - w^3 &= 326.2 \\ w^3 - x^3 &= 61. \end{aligned}$$

Required, to find the values of u, v, w, x, y, z .

The fact that the six quantities are the sides and diagonals of a quadrilateral provides one equation that they must satisfy, namely:

$$u^4 w^2 + x^2 (v^4 + w^2 y^2 - v^2 (w^2 - x^2 + y^2)) - (y^2 (w^2 + x^2 - y^2) + v^2 (-w^2 + x^2 + y^2)) z^2 + y^2 z^4 - u^2 (v^2 (w^2 + x^2 - y^2) + w^2 (-w^2 + x^2 + y^2) + (w^2 - x^2 + y^2) z^2) = 0.$$

This equation, together with the five given conditions, provides a complete set of equations for the six quantities. However, Seki Kōwa's explanation, which is only a sketch, does not mention this sixth equation, so it may be that what he solved was the indeterminate problem given by the other five equations. That, however, would be rather strange, since then the quadrilateral would play no role whatsoever in the problem. His solution is discussed in Sect. 3 of Chapter 14. Whatever the case, it is known that such equations were solved numerically by the Chinese using a counting board. Here once again it is very clear that the motive for the problem is algebraic, even though it does amount to a nontrivial investigation of the relations among the parts of a quadrilateral.

4.2. Beginnings of the calculus in Japan. By the end of the seventeenth century the *wasanists* were beginning to use techniques that resemble the infinitesimal methods being used in Europe about this time. Of course, in one sense Zu Chongzhi had used some principles of calculus 1000 years earlier in his application of Cavalieri's principle to find the volume of a sphere. The intuitive basis of the principle is that equals added to equals yield equal sums, and a solid can be thought of as the sum of its horizontal sections. It isn't really, of course. No finite sum of areas and no limit of such a sum can ever have positive volume. Students in calculus courses learn to compute volumes using approximating sums that are very thin prisms, but not infinitely thin.

In Japan these techniques were first applied in the area called *yenri* (circle theory),¹¹ a topic that had been studied extensively in China. The idea of approximating by shells or disks can be seen in the 1684 edition of the *Ketsugi-shō* (*Combination Book*), first published in 1660 by Isomura Kittoku.

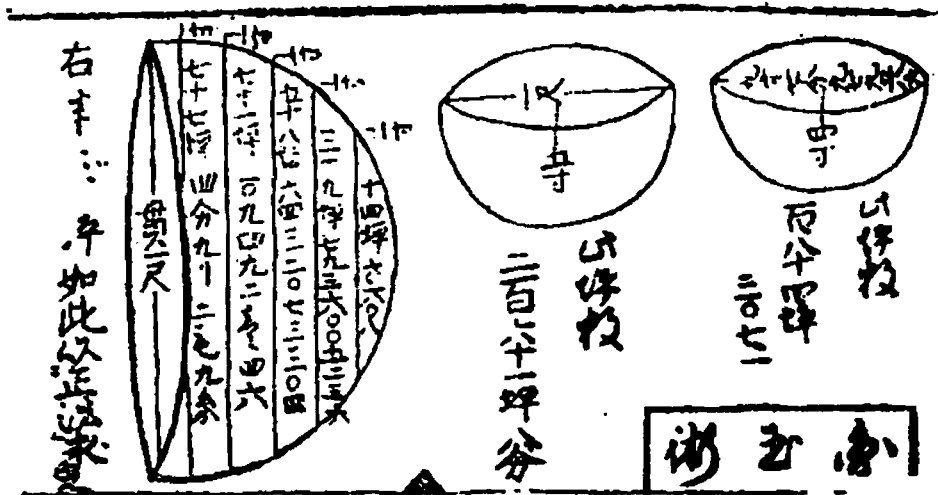
Isomura Kittoku explained the method as follows (Mikami, 1913, p. 204):

If we cut a sphere of diameter 1 foot into 10,000 slices, the thickness of each slice is 0.001 feet, which will be something like that of a very thin paper. Finding in this way the volume of each of them, we sum up the results, 10,000 in number, when we get 532.6 measures [that is, a volume of 0.5326 cubic foot]. Besides, it is true, there are small incommensurable parts, which are neglected.

The technique of obtaining extraordinary precision and using it to perform numerical experiments which provide the basis for general assertions also appears in some remarkable infinite series attributed to Takebe Kenkō, as we shall see below. Takebe Kenkō's method of squaring the circle was based on a relation, which he apparently discovered in 1722, between the square of half of an arc, the height h of the arc,¹² and the diameter d of the circle. Here is his own description of this discovery, as explained by Smith and Mikami (1914, pp. 147-149). He began with height $h = 0.000001 = 10^{-6}$ and $d = 10$, finding the square of the arc geometrically with accuracy to 53 decimal places.

The value of the square of this arc is

0.00001 00000 00333 33335 11111 12253 96833 52381 01394 90188 203+.



Isomura Kittoku's method of computing the volume of a sphere.

© Stock Montage, Inc.

According to Smith and Mikami (1914, p. 148), the value given by Takebe Kenkō was

¹¹ The symbol for circle here (*yen*) is also the symbol for the Japanese unit of currency; it is actually pronounced "en."

¹² This height is called the *sagitta* (arrow) by lens grinders, a name first bestowed on it in India. It is now called the *versed sine* in mathematics.

0.00000 00000 33333 35111 11225 39690 66667 28234 77694 79595 875+.

But this value does not fit with the procedure followed by Takebe Kenkō; it does not even yield the correct first approximation. The figure given by Smith and Mikami appears to represent the value obtained by Takebe Kenkō *after* the first approximation was subtracted, but with the result multiplied by the square of the diameter.¹³ In appreciating Takebe Kenkō's method, the first problem to be solved is the source of this extremely accurate measurement of the circle. According to Smith and Mikami (1914, p. 148), Takebe Kenkō said that the computation was given in two other works, both of which are now lost, leaving us to make our own conjectures. The first clue that strikes us in this connection is the seemingly strange choice of the *square* of the arc rather than the arc itself. Why would it be easier to compute the square of the arc than the arc itself? An answer readily comes to mind: The arc is approximated by its chord, and the chord is one side of a convenient right triangle. In fact, the chord is the mean proportional between the diameter of the circle and the height of the arc, so that in this case it is $\sqrt{dh} = \sqrt{10^{-5}}$. When we square it, we get just $dh = 10^{-5}$, which acts as Takebe Kenkō's first approximation. That result suggests that the length of the arc was reached by repeatedly bisecting the arc, taking the chord as an approximation. This hypothesis gains plausibility, since it is known that this technique had been used earlier to approximate π . Since $a^2 = 4(a/2)^2$, it was only necessary to find the square of half the arc, then multiply by 4. The ratio of the chord to the diameter is even easier to handle, especially since Takebe Kenkō has taken the diameter to be 10. If x is the square of this ratio for a given chord, the square of ratio for the chord of half of the arc is $(1 - \sqrt{1-x})/2$. In other words, the iterative process $x \mapsto (1 - \sqrt{1-x})/2$ makes the bisection easy. If we were dealing with the arc instead of its square, each step in that process would involve two square roots instead of one. Even as it is, Takebe Kenkō must have been a calculating genius to iterate this process enough times to get 53 decimal places of accuracy without making any errors. The result of 50 applications yields a ratio which, multiplied by $100 \cdot 4^{50}$, is

0.00001 00000 00333 33335 11111 12253 96833 52381 01131 94822 94294 362+.

This number of iterations gives 38 decimal places of accuracy. Even with this plausible method of procedure, it still strains credibility that Takebe Kenkō achieved the claimed precision. However, let us pass on to the rest of his method.

After the first approximation hd is subtracted, the new error is 10^{-12} times 0.3333333..., which suggests that the next correction should be $10^{-12}/3$. But this is exactly $h^2/3$, in other words $h/(3d)$ times the first term. When it is subtracted from the previously corrected value, the new error is

$10^{-19} \cdot 0.17777 77892 06350 01904 76806 15685 4870 +$.

The long string of 7's here suggests that this number is 10^{-19} times $\frac{1}{10} + \frac{7}{90} = \frac{16}{90} = \frac{8}{45}$, which is $(8h)/(15d)$ times the previous correction. By continuing for a few more terms, Takebe Kenkō was able to observe a pattern: The corrections are obtained by multiplying successively by $h/(3d)$, $(8h)/(15d)$, $(9h)/(14d)$, $(32h)/(45d)$, $(25h)/(33d)$,... Some sensitivity to the factorization of integers is necessary to

¹³ Even so, there is one 3 missing at the beginning and, after it is restored, the accuracy is "only" 33 decimal places. That precision, however, would have been all that Takebe Kenkō needed to compute the four corrections he claimed to have computed.

see the recursive operation: multiplication by $(h/d)[2n^2/(n+1)(2n+1)]$. Putting these corrections together as an infinite series leads to the expression

$$\frac{a^2}{4} = dh \left[1 + \sum_{n=1}^{\infty} \frac{2^{2n+1}(n!)^2}{(2n+2)!} \cdot \left(\frac{h}{d}\right)^n \right]$$

when the full arc has length a .

In using this numerical approach, Takebe Kenkō had reached his conclusion inductively. This induction was based on a faith (which turns out to be justified) that the coefficients of the power series are rational numbers that satisfy a fairly simple recursive formula. As you know, the power series for the sine, cosine, exponential, and logarithm have this happy property, but the series for the tangent, for example, does not.

This series solves the problem of rectification of the circle and hence all problems that depend on knowing the value of π . In modern terms the series given by Takebe Kenkō represents the function

$$\left(d \arcsin \left(\sqrt{\frac{h}{d}} \right) \right)^2.$$

Takebe Kenkō's discovery of this result in 1722 falls between the discovery of the power series for the arcsine function by Newton in 1676 and its publication by Euler in 1737.

Was European calculus transmitted to Japan in the seventeenth century? The methods used by Isomura Kittoku to compute the volume and surface area of a sphere and by Takebe Kenkō to compute the square of a half-arc in terms of the versed sine of the arc are at the heart of calculus. Smith and Mikami (1914, pp. 148–155) argue that some transmission from Europe at this time is plausible in the case of Takebe Kenkō. They note that there was some contact, although very limited, between Japanese and European scholars, even during the period of “closure,” and that a Jesuit missionary in China named Pierre Jartoux (1668–1721) communicated some of the latest European discoveries to his Chinese hosts. After noting that “there is no evidence that Seki or his school borrowed their methods from the West” (1914, p. 142), they argue as follows (1914, p. 155):

Here then is a scholar, Jartoux, in correspondence with Leibnitz [sic], giving a series not difficult of deduction by the calculus, which series Takebe uses and which is the essence of the *yenri*, but which Takebe has difficulty in explaining... [I]t seems a reasonable conjecture that Western learning was responsible for [Jartoux'] work, that he was responsible for Takebe's series, and that Takebe explained the series as best he could.

Probably the question of Western influence on Japanese mathematics cannot be decided. However, in allowing for the possibility of communication from West to East, we must not neglect the possibility of some transmission in the opposite direction, in addition to what was transmitted from the Muslims and Hindus earlier. Leibniz, in particular, was fascinated with oriental cultures, and at least two of his results, one of them a simple observation on determinants and the other a more extensive development of combinatorics, were known earlier in India and Japan. It should also be noted that in contrast to the Chinese mathematicians,

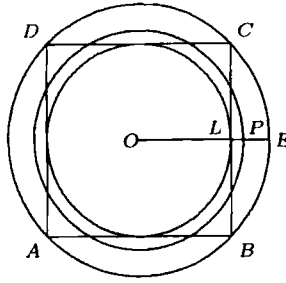


FIGURE 10. Rounding a square.

the practitioners of *wasan* were not immediately attracted to Euclid when his work arrived in Japan in the nineteenth century. According to Murata (1994, p. 109), having seen Chinese translations of Euclid, they were repelled by the great amount of fuss required to derive elementary facts. They may have taken just the ideas that appealed to them out of the information reaching them through contacts with the Chinese.

5. India

The *Sulva Sutras* contain many transformation-of-area constructions such as are later found in Euclid. In particular, the Pythagorean theorem, and constructions for finding the side of a square equal to a rectangle, or the sum or difference of two other squares are given. This construction resembles the one found in Proposition 5 of Book 2 of Euclid rather than Euclid's construction of the mean proportional in Book 6, both of which are discussed in Chapter 10. The Pythagorean theorem is not given a name, but is stated as the fact that "the diagonal of a rectangle produces both [areas] which its length and breadth produce separately." Among other transformation of area problems the Hindus considered in particular the problem of squaring the circle. The *Bodhayana Sutra* states the converse problem of constructing a circle equal to a given square. The construction is shown in Fig. 10, where $LP = \frac{1}{3}LE$.

In terms that we can appreciate, this construction gives a value for two-dimensional π of $18(3 - 2\sqrt{2})$, which is about 3.088.

5.1. Aryabhata I. Chapter 2 of Aryabhata's *Aryabhatiya* (Clark, 1930, pp. 21–50) is called *Ganitapada* (*Mathematics*). In Stanza 6 of this chapter Aryabhata gives the correct rule for area of a triangle, but declares that the volume of a tetrahedron is half the product of the height and the area of the base. He says in Stanza 7 that the area of a circle is half the diameter times half the circumference, which is correct, and shows that he knew that one- and two-dimensional π were the same number. But he goes on to say that the volume of a sphere is the area of a great circle times its own square root. This would be correct only if three-dimensional π equaled $\frac{16}{9}$, very far from the truth! Yet Aryabhata knew a very good approximation to one-dimensional π . In Stanza 10 he writes:

Add 4 to 100, multiply by 8, and add 62,000. The result is approximately the circumference of a circle of which the diameter is 20,000.

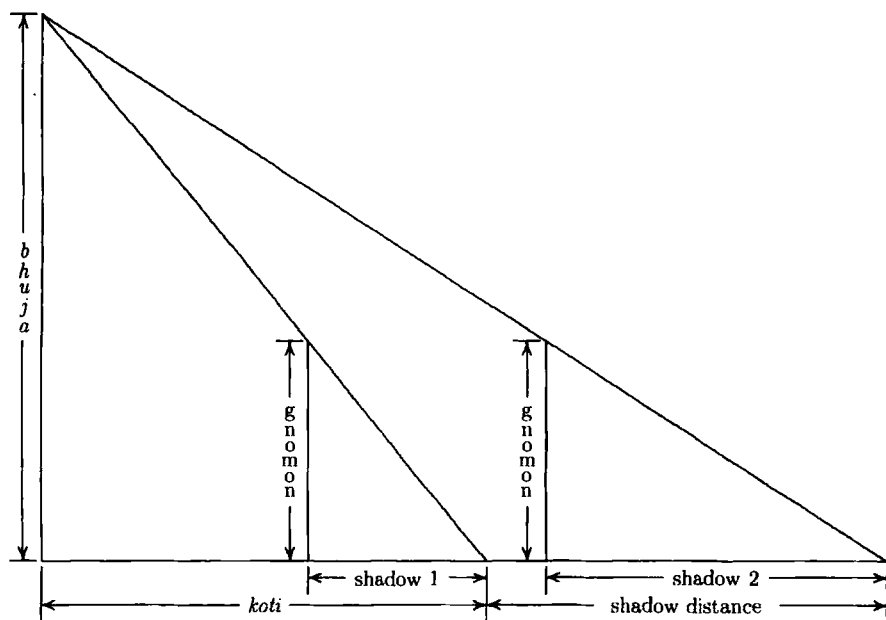


FIGURE 11. The Hindu variant of the double-difference method of surveying.

This procedure gives a value of one-dimensional π equal to 3.1416, which is quite accurate indeed. It exceeds the true value by less than 0.01%.

Aryabhata also knows about the double-difference method of surveying that we discussed above. Whether this knowledge is a case of transmission or independent discovery is not clear. The rule given is slightly different from the discussion that accompanies Fig. 6 and is illustrated by Fig. 11.

The distance between the ends of the two shadows multiplied by the length of the shadow and divided by the difference in length of the two shadows give the *koti*. The *koti* multiplied by the length of the gnomon and divided by the length of the shadow gives the length of the *bhuja*. [Clark, 1930, p. 32]

Trigonometry. The inclusion of this variant of the double-difference method of surveying in the *Aryabhatiya* presents us with a small puzzle. As a method of surveying, it is not efficient. It would seem to make more sense to measure angles rather than using only right angles and measuring many more lines. But angles are really not involved here. It is possible to have a clear picture of two mutually perpendicular lines without thinking "right angle." The notion of angles in general as a species of mathematical objects—the figures formed by intersecting lines, which can be measured, added, and subtracted—appears to be a Greek innovation in the sixth and fifth centuries BCE, and it seems to occur only in plane geometry. Its origins may be in stonemasonry and carpentry, where regular polygons have to be fitted together. Astronomy probably also made some contribution.

The earliest form of trigonometry that we can recognize was a table of correspondences between arcs and their chords. We know exactly how such a table was originally constructed, since an explanation can be found in Ptolemy's treatise on astronomy, written around 150 CE.

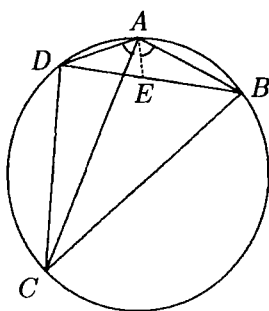


FIGURE 12. For a quadrilateral inscribed in a circle, the product of the diagonals equals the sum of the products of the two pairs of opposite sides.

To construct his table of chords, Ptolemy had to make use of some subtle geometry developed earlier: in particular, the fact that for a quadrilateral inscribed in a circle the product of the diagonals is the sum of the products of the two pairs of opposite sides.¹⁴ Ptolemy proved this result by drawing AE (see Fig. 12) so that $\angle BAE = \angle DAC$, thus obtaining two pairs of similar triangles: $\triangle BAE \sim \triangle CAD$ and $\triangle ADE \sim \triangle ACB$. (Angles ABD and ACD are equal, both being inscribed in the same arc \widehat{AD} ; and similarly $\angle ACB = \angle ADB$.) Ptolemy used this relation to compute the chord of the difference of two arcs and the chord of half an arc.

Since Ptolemy knew the construction of the regular dodecagon and the regular decagon, he was easily able to compute the chords of 36° and 30° , expressed in units of one-sixtieth of the radius. His difference theorem then gave the chord of 6° . Then by repeated bisection he got the chord of 3° , then $1^\circ 30'$, and finally, $45'$. Using these two values and certain inequalities, he was able to set upper and lower bounds on the chord of 1° with sufficient precision for his purposes. He then set out a table with 360 entries, giving the chords of arcs at half-degree increments up to 180° .

Although this table fulfilled its purpose in astronomy, the chord is a cumbersome tool to use in studying plane geometry. For example, it was well known that in any triangle, the angle opposite the larger of two sides will be larger than the angle opposite the smaller side. But what is the exact, quantitative relation between the two sides and the two angles? The ratio of the sides has no simple relationship to the ratio of the angles or to the chords of those angles. There is, however, a very simple relation between the sides and the chords of *twice* the opposite angles, that is, the chords these angles cut off on the circumscribed circle. One might have thought that the constant comparison of a chord with the diameter would have inspired someone to associate the arc with the angle inscribed in it rather than the central angle it subtends. After all, a side of any triangle is the chord of a central angle in the circumscribed circle equal to the double of the opposite angle. Hindu astronomers discovered that trigonometry is simpler if you express the relations between circular arcs and chords in terms of half-chords, what are now called *sines*.

¹⁴ When the quadrilateral is a rectangle, this fact is the form of the Pythagorean theorem given in the *Sulva Sūtras*. Gow (1884, p. 194) describes this result as "now appended to Euclid VI."

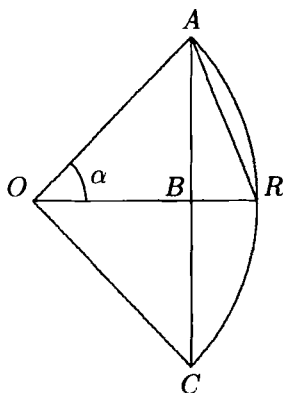


FIGURE 13. The “bowstring” diagram. The sine of the arc \widehat{AR} is the line AB .

In Fig. 13 the arc \widehat{AR} can be measured by either line AB or AR . Ptolemy chose AR and was led to the complications already mentioned. The Hindus preferred AB , which is succinctly described as half the chord of twice the arc. We mentioned above that the Chinese word (*xian*) for the hypotenuse of a right triangle means *bowstring*. The Hindus used the Sanskrit term for a bowstring (*jya* or *jiva*) to mean the sine. The reason for the colorful language is obvious from the figure.

To all appearances, then, trigonometry began to assume its modern form among the Hindus some 1500 years ago. A few reservations are needed, however. First, for the Hindu mathematicians the *sine* was not, as it is to us, a *ratio*. It was a *length*, and that physical dimension had to be taken into account in all computations. Second, the only Hindu concept corresponding approximately to our trigonometric functions was the sine. The tangent, secant, cosine, cotangent, and cosecant were not used. Third, the use of trigonometry was restricted to astronomy. Surveying, which is the other natural place to use trigonometry, did not depend on angle measurement.

Aryabhata used the sine function developed in the *Surya Siddhanta*, giving a table for computing its values at intervals of 225' ($3^\circ 45'$) of arc from 0° to 90° degrees and expressing these values in units of 1' of arc, rounded to the nearest integer, so that the sine of 90° is $3438 = 360 \cdot 60 \cdot \pi$. This interval suggests that the tables were computed independently of Ptolemy's work. If the Hindu astronomers had read Ptolemy, their tables of sines could easily have been constructed from his table of chords, and with more precision than is actually found. Almost certainly, this interval was reached by starting with an angle of 30° , whose sine was known to be half of the radius, then applying the formula for a half-angle to get successively 15° , $7^\circ 30'$, and finally $3^\circ 45'$. Aryabhata's table is actually a list of the *differences* of 24 successive sines at intervals of 225 minutes. Since one minute of arc is a very small quantity relative to the radius, the 24 values provide sufficient precision for the observational technology available at the time. Notice, however, that to calculate the sine of half of an angle θ one would have to apply the cumbersome

formula

$$\sin \frac{\theta}{2} = \sqrt{\frac{3438 - \sqrt{3438^2 - \sin^2 \theta}}{2}}.$$

We can therefore well understand why Aryabhata did not refine his table further. Aryabhata's list of sine differences is the following:

225, 224, 222, 219, 215, 210, 205, 199, 191, 183, 174,
164, 154, 143, 131, 119, 106, 93, 79, 65, 51, 37, 22, 7.

A comparison with a computer-generated table for the same differences reveals that Aryabhata's table is accurate except that his sixth entry should be 211 and the eighth should be 198. But surely an error of less than half of 1% is not a practical matter, and Aryabhata definitely took the practical approach. He explained that his table of sine differences was computed by a recursive procedure, which can be described in our terms as follows (Clark, 1930, p. 29). Starting with $d_1 = 225$,

$$d_{n+1} = d_n - \frac{d_1 + \cdots + d_n}{d_1},$$

where each term is rounded to the nearest integer after being calculated from this formula.

Aryabhata applied the sine function to determine the elevation of the Sun at a given hour of the day. The procedure is illustrated in Fig. 14 for an observer located at O in the northern hemisphere on a day in spring or summer. This figure shows a portion of the celestial sphere. The arc $RETSWV$ is the portion of the great circle in which the observer's horizontal plane intersects the sphere. The Sun will rise for this observer at the point R and set at the point V . The arc is slightly larger than a semicircle, since we are assuming a day in spring or summer. The chord RV runs from east to west. The Sun will move along the small circle RHV at a uniform rate, and the plane of this circle is parallel to the equatorial circle EMW . (At the equinox, the "day-circle" RV coincides with the equatorial circle EW .) Aryabhata gave the correct formula for finding the radius of this day-circle in terms of the elevation of the Sun above the celestial equator and the radius of the celestial sphere. That radius is the sine of the co-declination of the Sun. Although Aryabhata had the concept of co-latitude, which served him in places where we would use the cosine function, for some reason he did not use the analogous concept of co-declination. As a result, he had to subtract the square of the sine of the declination from the square of the radius of the celestial sphere, then take the square root.

The point Z is the observer's zenith, M is the point on the celestial equator that is due south to the observer, and S is the point due south on the horizon, so that the arc \widehat{ZM} is the observer's latitude, and the two arcs \widehat{ZN} and \widehat{MS} are both equal to the observer's co-latitude. The point H is the location of the Sun at a given time, MF and HG are the projections of M and H respectively on the horizontal plane, and HK is the projection of H on the chord RV . Finally, the great-circle arc HT , which runs through Z , is the elevation of the Sun. The problem is to determine its sine HG in terms of lengths that can be measured.

Because their sides are parallel lines, the triangles MOF and HKG are similar, so that $MO : HK = MF : HG$. Hence we get

$$HG = \frac{HK \cdot MF}{MO}.$$

the square root of the product is the area. In our terms this rule says that the area of a quadrilateral of sides a , b , c , and d is $\sqrt{(s-a)(s-b)(s-c)(s-d)}$, where s is half of the sum of the lengths of the sides. The case when $d = 0$, which is a triangle, is known as *Heron's formula*. Brahmagupta did not mention the restriction that the quadrilateral must be a cyclic quadrilateral, that is, it must be inscribed in a circle.

Like Aryabhata, Brahmagupta knew that what we are calling one- and two-dimensional π were the same number. In Stanza 40 he says that when the diameter and the square of the radius respectively are multiplied by 3, the results are the "practical" circumference and area. In other words, $\pi = 3$ is a "practical" value. He also gives the "neat" ("exact") value as $\sqrt{10}$. Since $\sqrt{10} = 3.1623$, this value is not an improvement on Aryabhata's 3.1416 in terms of accuracy. If one had to work with π^2 , however, it might be more convenient. But π^2 occurs in very few contexts in mathematics, and none at all in elementary mathematics.

Section 5 of Chapter 12 of the *Brahmasphutasiddhanta* gives a rule for finding the volume of a frustum of a rectangular pyramid. In keeping with his approach of giving approximate rules, Brahmagupta says to take the product of the averages of the sides of the top and bottom in the two directions, then multiply by the depth. He calls this result the "practical measure" of the volume, and he knew that this simple rule gave a volume that was too small.

For his second approximation, which he called the "rough" volume, he took the average of the areas of the top and bottom and multiplied by the depth.¹⁵ He also knew that this procedure gave a volume that was too large. The actual volume lies between the "practical" volume and the "rough" volume, but where? We know that the actual volume is obtained as a mixture of two parts "practical" and one part "rough", and so did Brahmagupta. His corrective procedure to give the "neat" (exact) volume was: Subtract the practical from the rough, divide the difference by three, then add the quotient to the practical value.

The phrasing of this result cries out for speculation on its origin. Why use the "practical" volume twice? Why not simply say, "The exact volume is two-thirds of the practical volume plus one-third of the rough volume"? Surely Brahmagupta could do this computation as well as we can and could have used this simpler language. Perhaps his roundabout way of expressing the result reveals the analysis by which he discovered it. Let us investigate what happens when we subtract the "practical" volume from the "rough" volume. First of all, since each is merely an area times the height of the frustum, we are really just subtracting the average area of two rectangles from the area of the rectangle formed by the averages of their parallel sides. Let us simplify by taking the case of two squares of sides a and b . What we are getting, then, is the average of the squares minus the square of the average:

$$\frac{a^2 + b^2}{2} - \left(\frac{a + b}{2}\right)^2.$$

Figure 15 shows immediately that this difference is just the square on side $(a - b)/2$. In that figure, half of the squares of sides a and b are set down with their diagonals in a straight line. The two isosceles right triangles below and to the right of the dashed lines fit together to form a square of side $(a + b)/2$. If the

¹⁵ This is the same procedure followed in the cuneiform tablet BM 85 194, discussed above in Subsection 2.3.

rectangle that is shaded dark, which lies inside these two isosceles triangles but outside the squares of sides b and a , is moved inside the square of side a so as to cover the rectangle that is shaded light, we see that the two isosceles triangles cover all of the two half-squares except for a square of side $(a - b)/2$. Since this figure is a very simple one, it seems likely that Brahmagupta would have known that the difference between his two estimates of the volume of a square frustum amounted to the volume of a prism of square base $(a - b)/2$ and height h .

But how did he know that he needed to take one-third of this prism, that is, the volume of a pyramid of the same base and height, and add it to the practical volume? To answer that question, consider a slight variant of the dissection shown in Fig. 3. First remove the four pyramids in the corners, each of which has volume $(h/3)((a - b)/2)^2$, which is exactly one-third of the difference between the gross and practical volumes. Doing so leaves a square platform with four "ramps" running down its sides. In our previous dissection we sliced off two of these ramps on opposite sides and glued them upside down on the other two ramps to make a "slab" of dimensions $a \times b \times h$. This time we slice off the outer half of all four ramps and bend them up to cover their upper halves. The result, shown in Fig. 16, is the cross-shaped prism of height h whose base is a square of side $(a + b)/2$ having a square indentation of side $\frac{a-b}{4}$ at each corner. Filling in these square prisms produces the volume that Brahmagupta called the practical measure. The volume needed to do so is $4h((a - b)/4)^2 = h((a - b)/2)^2$. Now three of the four pyramids removed from the corners, taken together, have exactly this much volume. If we use these three to fill in the practical volume, we have one pyramid left over, and its volume is one-third of the difference between the rough and practical volumes. A person who followed the dissection outlined above would then very naturally describe the volume of the pyramid as the practical volume plus one-third of the difference between the gross and practical volumes. That would be natural, but it would be rash to infer that Brahmagupta *did* imagine this dissection; all we have shown is that he *might have done* that.

Questions and problems

9.1. Show how it is possible to square the circle using ruler and compass given the assumption that $\pi = (16\sqrt{2})/7$.

9.2. Prove that the implied Egyptian formula for the volume of a frustum of a square pyramid is correct. If the sides of the upper and lower squares are a and b and the height is h , the implied formula is:

$$V = \frac{h}{3}(a^2 + ab + b^2).$$

9.3. Looking at the Egyptian pyramids, with their layers of brick revealed, now that most of the marble facing that was originally present has been removed, one can see that the total number of bricks must be $1 + 4 + 9 + \cdots + n^2$ if the slope (*seked*) is constant. Assuming that the Egyptian engineers had the kind of numerical knowledge that would enable them to find this sum as $\frac{1}{6}n(n + 1)(2n + 1)$, can you conjecture how they may have arrived at the formula for the volume of a frustum? Is it significant that in the only example we have for this computation, the height is 6 units?

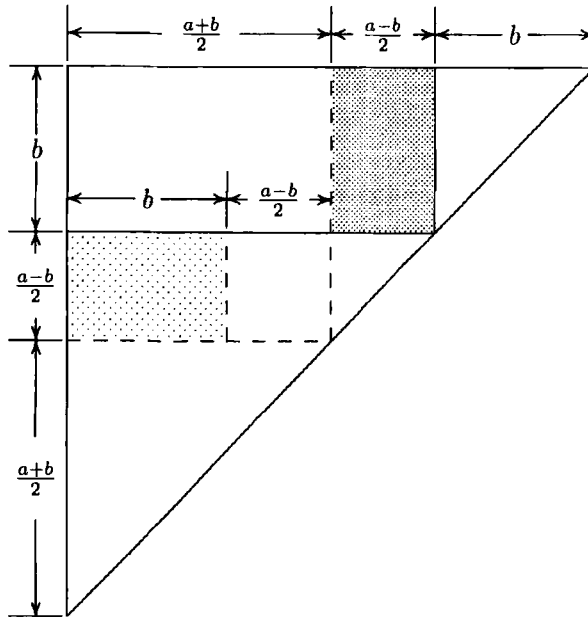


FIGURE 15. The average of the squares minus the square of the average is the square of the semi-difference.

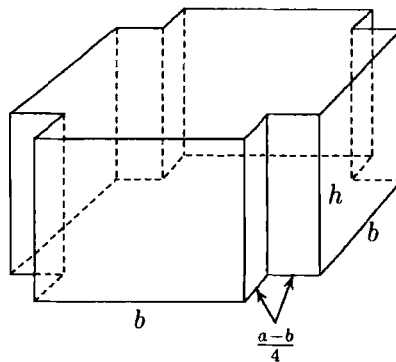


FIGURE 16. Frustum of a pyramid with its corners removed and side "ramps" folded up to form a cross-shaped prism.

9.4. Explain the author's solution of the following problem from the cuneiform tablet BM 85 196. Here the numbers in square brackets were worn off the tablet and have been reconstructed.

A beam of length 0;30 GAR is leaning against a wall. Its upper end is 0;6 GAR lower than it would be if it were perfectly upright. How far is its lower end from the wall?

Do the following: Square 0;30, obtaining 0;15. Subtracting 0;6 from 0;30 leaves 0;24. Square 0;24, obtaining 0;9,36. Subtract

0;9,36 from $[0;15]$, leaving 0;5,24. What is the square root of 0;5,24? The lower end of the beam is $[0;18]$ from the wall.

When the lower end is 0;18 from the wall, how far has the top slid down? Square 0;18, obtaining 0;5,24....

9.5. Show that the average of the areas of the two bases of a frustum of a square pyramid is the sum of the squares of the average and semidifference of the sides of the bases. Could this fact have led the Mesopotamian mathematicians astray in their computation of the volume of the frustum? Could the analogy with the area of a trapezoid have been another piece of misleading evidence pointing toward the wrong conclusion?

9.6. The author of the *Zhou Bi Suan Jing* had a numerical method of finding the length of the diagonal of a rectangle of width a and length b , which can be described as follows. Square the sum of width and length, subtract twice the area, then take the square root. Should one conclude from this that the author knew that the square on the hypotenuse was the sum of the squares on the legs?

9.7. What happens to the estimate of the Sun's altitude (36,000 km) given by Zhao Shuang if the "corrected" figure for shadow lengthening (4 *fen* per 1000 *li*) is used in place of the figure of 1 *fen* per 1000 *li*?

9.8. The *gougu* section of the *Jiu Zhang Suanshu* contains the following problem:

Under a tree 20 feet high and 3 in circumference there grows a vine, which winds seven times the stem of the tree and just reaches its top. How long is the vine?

Solve this problem.

9.9. Another right-triangle problem from the *Jiu Zhang Suanshu* is the following. "There is a string hanging down from the top of a pole, and the last 3 feet of string are lying flat on the ground. When the string is stretched, it reaches a point 8 feet from the pole. How long is the string?" Solve this problem. You can also, of course, figure out how high the pole is from this information.

9.10. A frequently reprinted problem from the *Jiu Zhang Suanshu* is the "broken bamboo" problem: A bamboo 10 feet high is broken and the top touches the ground at a point 3 feet from the stem. What is the height of the break? Solve this problem, which reappeared several centuries later in the writings of the Hindu mathematician Brahmagupta.

9.11. The *Jiu Zhang Suanshu* implies that the diameter of a sphere is proportional to the cube root of its volume. Since this fact is equivalent to saying that the volume is proportional to the cube of the diameter, should we infer that the author knew both proportions? More generally, if an author knows (or has proved) "fact A," and fact A is logically equivalent to fact B, is it accurate to say that the author knew or proved fact B? (See also Problem 9.6 above.)

9.12. Show that the solution to the quadrilateral problem of Sawaguchi Kazuyuki is $u = 9$, $v = 8$, $w = 5$, $x = 4$, $y = \sqrt{(1213 + 69\sqrt{273})/40}$, $z = 10$. (The approximate value of y is 7.6698551.) From this result, explain how Sawaguchi Kazuyuki must have invented the problem and what the two values 60.8 and 326.2

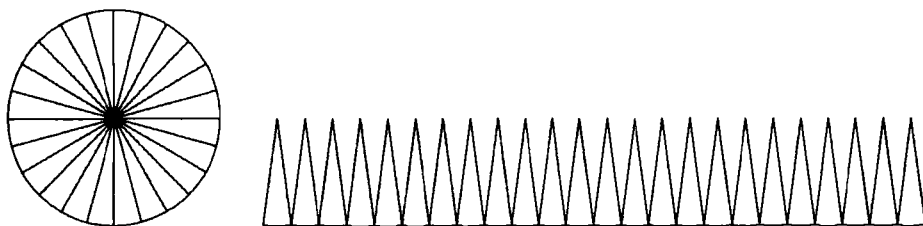


FIGURE 17. A disk cut into sectors and opened up.

are approximations for. How does this problem illustrate the claim that these challenge problems were algebraic rather than geometric?

9.13. How is it possible that some Japanese mathematicians believed the area of the sphere to be one-fourth the square of the circumference, that is, $\pi^2 r^2$ rather than the true value $4\pi r^2$? Smith and Mikami (1914, p. 75) suggest a way in which this belief might have appeared plausible. To explain it, we first need to see an example in which the same line of reasoning really does work.

By imagining a circle sliced like a pie into a very large number of very thin pieces, one can imagine it cut open and all the pieces laid out next to one another, as shown in Fig. 17. Because these pieces are very thin, their bases are such short segments of the circle that each base resembles a straight line. Neglecting a very tiny error, we can say that if there are n pieces, the base of each piece is a straight line of length $2\pi r/n$. The segments are then essentially triangles of height r (because of their thinness), and hence area $(1/2) \cdot (2\pi r^2)/n$. Since there are n of them, the total area is πr^2 . This heuristic argument gives the correct result. In fact, this very figure appears in a Japanese work from 1698 (Smith and Mikami, 1914, p. 131).

Now imagine a hemispherical bowl covering the pie. If the slices are extended upward so as to slice the bowl into equally thin segments, and those segments are then straightened out and arranged like the segments of the pie, they also will have bases equal to $\frac{2\pi r}{n}$, but their height will be one-fourth of the circumference, in other words, $\pi r/2$, giving a total area for the hemisphere of $(1/2) \cdot \pi^2 r^2$. Since the area is $2\pi r^2$, this would imply that $\pi = 4$. What is wrong with the argument? How much error would there be in taking $\pi = 4$?

9.14. What is the justification for the statement by the historian of mathematics T. Murata that Japanese mathematics (*wasan*) was not a science but an art?

9.15. Show that Aryabhata's list of sine differences can be interpreted in our language as the table whose n th entry is

$$3438 \left[\sin \left(\frac{n\pi}{48} \right) - \sin \left(\frac{(n-1)\pi}{48} \right) \right].$$

Use a computer to generate this table for $n = 1, \dots, 24$, and compare the result with Aryabhata's table.

9.16. If the recursive procedure described by Aryabhata is followed faithfully (as a computer can do), the result is the following sequence.

225, 224, 222, 219, 215, 210, 204, 197, 189, 181, 172,
162, 151, 140, 128, 115, 102, 88, 74, 60, 45, 30, 15, 0

Compare this list with Aryabhata's list, and note the systematic divergence. These differences should be approximately 225 times the cosine of the appropriate angle. That is, $d_n \approx 225 \cdot \cos(225(n + 0.5) \text{ minutes})$. What does that fact suggest about the source of the systematic errors in the recursive procedure described by Aryabhata?

9.17. Use Aryabhata's procedure to compute the altitude of the Sun above the horizon in London (latitude $51^\circ 32'$) at 10:00 AM on the vernal equinox. Assume that the sun rises at 6:00 AM on that day and sets at 6:00 PM.

9.18. Why is it necessary that a quadrilateral be inscribed in a circle in order to compute its diagonals knowing the lengths of its sides? Why is it not possible to do so in general?

9.19. Show that the formula given by Brahmagupta for the area of a quadrilateral is correct if and only if the quadrilateral can be inscribed in a circle.

9.20. Imagine a sphere as a polyhedron having a large number of very small faces. Deduce the relation between the volume of a sphere and its area by considering the pyramids obtained by joining the points of each face to the center of the sphere.

CHAPTER 10

Euclidean Geometry

We shall divide the history of Greek mathematics into four periods. The first period, from about 600 to 400 BCE, was the time when the Greeks acquired geometry from Egypt and Mesopotamia and turned it in the direction of logical argument. The second period came in the fourth century, when the logical aspects of the subject were debated in Plato's Academy¹ and Aristotle's Lyceum,² proofs were improved, and basic principles and assumptions were isolated. The third period began in the third century, when the mature subject was expounded in Euclid's *Elements*, and further research continued on more complicated curves and surfaces. The fourth and final period was a long decline in originality, in which no revolutionary changes occurred and commentaries were the main literary form.

1. The earliest Greek geometry

The history of Greek mathematics up to the time of Euclid (300 BCE) was written by Eudemus, a pupil of Aristotle. This history was lost, but it is believed to be the basis of the first paragraph of a survey given by Proclus in the fifth century CE in the course of his commentary on the first book of Euclid. In this passage Proclus mentions 25 men who were considered to have made significant contributions to mathematics. Of these 25, five are well known as philosophers (Thales, Pythagoras, Anaxagoras, Plato, and Aristotle); three are famous primarily as mathematicians and astronomers (Euclid, Eratosthenes, and Archimedes). The other 17 have enjoyed much less posthumous fame. Some of them are so obscure that no mention of them can be found anywhere except in Proclus' summary. Some others (Theodorus, Archytas, Menaechmus, Theaetetus, and Eudoxus) are mentioned by other commentators and by Plato. The 13 just named are the main figures we shall use to sketch the history of Greek geometry. It is clear from what Proclus writes that something important happened to mathematics during the century of Plato and Aristotle, and the result was a unique book, Euclid's *Elements*.

Missing from the survey of Proclus is any reference to Mesopotamian influence on Greek geometry. This influence is shown clearly in Greek astronomy, in the

¹ This word has become so common in English that its original, legendary meaning is mostly forgotten. In his biography of the Athenian king Theseus, who had slain the Minotaur on Crete as a youth, Plutarch says that at the age of 50 the widowed king abducted the beautiful 12-year-old Helen of Sparta and hid her away. (This was before she married Menelaus and ran off with Paris, becoming the cause of the Trojan War.) Her twin brothers Castor and Polydeukes (Pollux) threatened to destroy Athens in revenge. Akademos, however, averted the calamity by telling them where she was hidden. For this deed he was venerated as the savior of the city, and a grove of trees on its northwest side, supposedly his burial place, was dedicated to his memory. Plato gave his lectures in that grove, and hence arose the phrase "the groves of Academe."

² Here is another word whose origins are lost in common usage. The Lyceum was so named because it was near the temple to Apollo Lykeios ("Apollo of the Wolves").

use of the sexagesimal system of measuring angles and in Ptolemy's explicit use of Mesopotamian astronomical observations. It *may* also appear in Book 2 of Euclid's *Elements*, which contains geometric constructions equivalent to certain algebraic relations that are frequently encountered in the cuneiform tablets. This relation, however, is controversial. Leaving aside the question of Mesopotamian influence, we do see a recognition of their debt to Egypt, which the Greeks never concealed. And how could they? Euclid actually lived in Egypt, and the other two of the "big three" Greek geometers, Archimedes and Apollonius, both studied there, in the Hellenistic city of Alexandria at the mouth of the Nile.

1.1. Thales. The philosopher Thales, who lived in the early sixth century BCE, was a citizen of Miletus, a Greek colony on the coast of Asia Minor. The ruins of Miletus are now administered by Turkey. Herodotus mentions Thales in several places. Discussing the war between the Medes and the Lydian king Croesus, which had taken place in the previous century, he says that an eclipse of the Sun frightened the combatants into making peace. Thales, according to Herodotus, had predicted that an eclipse would occur no later than the year in which it actually occurred. Herodotus goes on to say that Thales had helped Croesus to divert the river Halys so that his army could cross it.

These anecdotes show that Thales had both scientific and practical interests. His prediction of a solar eclipse, which, according to the astronomers, occurred in 585 BCE, seems quite remarkable, even if, as Herodotus says, he gave only a period of several years in which the eclipse was to occur. Although solar eclipses occur regularly, they are visible only over small portions of the Earth, so that their regularity is difficult to discover and verify. Lunar eclipses exhibit the same period as solar eclipses and are easier to observe. Eclipses recur in cycles of about 19 solar years, a period that seems to have been known to many ancient peoples. Among the cuneiform tablets from Mesopotamia there are many that discuss astronomy, and Ptolemy uses Mesopotamian observations in his system of astronomy. Thales could have acquired this knowledge, along with certain simple facts about geometry, such as the fact that the base angles of an isosceles triangle are equal. Bychkov (2001) argues that the recognition that the base angles of an isosceles triangle are equal probably did come from Egypt. In construction, for example, putting a roof on a house, it is not crucial that the cross section be exactly an isosceles triangle, since it is the ridge of the roof that must fit precisely, not the edges. However, when building a symmetric square pyramid, errors in the base angles of the faces would make it impossible for the faces to fit together tightly. Therefore, he believes, Thales must have derived this theorem from his travels in Egypt.

In his *Discourses on the Seven Wise Men*, Plutarch reports that Thales traveled to Egypt and was able to calculate the height of the Great Pyramid by driving a pole into the ground and observing that the ratio of the height of the pyramid to that of the pole was the same as the ratio of their shadow lengths. In his *Lives of Eminent Philosophers*, Diogenes Laertius cites the historian Hieronymus (fourth or third century BCE) in saying that Thales calculated the height of the pyramid by waiting until his shadow was exactly as long as he was tall, then measuring the length of the shadow of the Great Pyramid.³ There are practical difficulties in executing this

³ A very interesting mystery/historical novel by Denis Guedj, called *Le théorème du perroquet*, uses this history to connect its story line. An English translation of this novel now exists, *The Parrot's Theorem*, St. Martin's Press, New York, 2002.

plan, since one could not get into the Pyramid to measure the distance from the center to the tip of the shadow directly. One might use the Pythagorean theorem, which Thales could well have known, to measure the distance from the center of the pyramid to the point where its outer wall intersects the vertical plane through the top of the pyramid and the tip of its shadow. A simpler way of computing the distance, however, is to reflect a triangle about one of its vertices. This technique is known to have been used by Roman surveyors to measure the distance across a river without leaving shore.

According to Diogenes Laertius, a Roman historian named Pamphila, who lived in the time of Nero, credits Thales with being the first to inscribe a right triangle in a circle. To achieve this construction, one would have to know that the hypotenuse of the inscribed triangle is a diameter. Diogenes Laertius goes on to say that others attribute this construction to Pythagoras.

1.2. Pythagoras and the Pythagoreans. Half a century later than Thales the philosopher Pythagoras was born on the island of Samos, another of the Greek colonies in Ionia. No books of Pythagoras survive, but many later writers mention him, including Aristotle. Diogenes Laertius devotes a full chapter to the life of Pythagoras. He acquired even more legends than Thales. According to Diogenes Laertius, who cites the logicist Apollodorus, Pythagoras sacrificed 100 oxen when he discovered the theorem that now bears his name. If the stories about Pythagoras can be believed, he, like Thales, traveled widely, to Egypt and Mesopotamia. He gathered about him a large school of followers, who observed a mystical discipline and devoted themselves to contemplation. They lived in at least two places in Italy, first at Croton, then, after being driven out,⁴ at Metapontion, where he died sometime around 500 BCE.

According to Book I, Chapter 9 of *Attic Nights*, by the Roman writer Aulus Gellius (ca. 130–180), the Pythagoreans first looked over potential recruits for physical signs of being educable. Those they accepted were first classified as *akoustikoi* (auditors) and were compelled to listen without speaking. After making sufficient progress, they were promoted to *mathēmatikoi* (learners).⁵ Finally, after passing through that state they became *physikoi* (natural philosophers). In his *Life of Pythagoras* Iamblichus uses these terms to denote the successors of Pythagoras, who split into two groups, the *akoustikoi* and the *mathēmatikoi*. According to Iamblichus, the *mathēmatikoi* recognized the *akoustikoi* as genuine Pythagoreans, but the sentiment was not reciprocated. The *akoustikoi* kept the pure Pythagorean doctrine and regarded the *mathēmatikoi* as followers of the disgraced Hippasus mentioned in Chapter 8.

Diogenes Laertius quotes the philosopher Alexander Polyhistor (ca. 105–35 BCE) as saying that the Pythagoreans generated the world from *monads* (units). By adding a single monad to itself, they generated the natural numbers. By allowing the monad to move, they generated a line, then by further motion the line generated plane figures (polygons), the plane figures then moved to generate solids (polyhedra). From the regular polyhedra they generated the four elements of earth, air, fire, and water.

⁴ Like modern cults, the Pythagoreans attracted young people, to the despair of their parents. Accepting new members from among the local youth probably aroused the wrath of the citizenry.

⁵ Gellius remarks at this point that the word *mathēmatikoi* was being inappropriately used in popular speech to denote a “Chaldean” (astrologer).

1.3. Pythagorean geometry. Euclid's geometry is an elaboration and systematization of the geometry that came from the Pythagoreans via Plato and Aristotle. From Proclus and other later authors we have a glimpse of a fairly sophisticated Pythagorean geometry, intertwined with mysticism. For example, Proclus reports that the Pythagoreans regarded the right angle as ethically and aesthetically superior to acute and obtuse angles, since it was "upright, uninclined to evil, and inflexible." Right angles, he says, were referred to the "immaculate essences," while the obtuse and acute angles were assigned to divinities responsible for changes in things. The Pythagoreans had a bias in favor of the eternal over the changeable, and they placed the right angle among the eternal things, since unlike acute and obtuse angles, it cannot change without losing its character. In taking this view, Proclus is being a strict Platonist; for Plato's ideal Forms were defined precisely by their absoluteness; they were incapable of undergoing any change without losing their identity.

Proclus mentions two topics of geometry as being Pythagorean in origin. One is the theorem that the sum of the angles of a triangle is two right angles (Book 1, Proposition 32). Since this statement is equivalent to Euclid's parallel postulate, it is not clear what the discovery amounted to or how it was made.

The other topic mentioned by Proclus is a portion of Euclid's Book 6 that is not generally taught any more, called application of areas. However, that topic had to be preceded by the simpler topic of transformation of areas. In his *Nine Symposium Books*⁶ Plutarch called the transformation of areas "one of the most geometrical" problems. He thought solving it was a greater achievement than discovering the Pythagorean theorem and said that Pythagoras was led to make a sacrifice when he solved the problem. The basic idea is to convert a figure having one shape to another shape while preserving its area, as in Fig. 1. To describe the problem in a different way: Given two geometric figures A and B , construct a third figure C the same size as A and the same shape as B . One can imagine many reasons why this problem would be attractive. If one could find, for example, a square equal to any given figure, then comparing sizes would be simple, merely a matter of converting all areas into squares and comparing the lengths of their sides. But why stop at that point? Why not do as the Pythagoreans apparently did, and consider the general problem of converting any shape into any other? For polygons this problem was solved very early, and the solution appears in very elegant form as Proposition 25 of Euclid's Book 6.

Related to the transformation of areas is the problem of application of areas. There are two such problems, both involving a given straight line segment AB and a planar polygon Γ . The first problem is to construct a parallelogram equal to Γ on part of the line segment AB in such a way that the parallelogram needed to fill up a parallelogram on the entire base, called the *defect*, will have a prescribed shape. This is the problem of *application with defect*, and the solution is given in Proposition 28 of Book 6. The second application problem is to construct a parallelogram equal to Γ on a base containing the line AB and such that the portion of the parallelogram extending beyond AB (the *excess*) will have a prescribed shape. This is the problem of *application with excess*, and the solution is Proposition 29 of Book 6. The construction for application with defect is shown in Fig. 2. This

⁶ The book is commonly known as *Convivial Questions*. The Greek word *sympósiōn* means literally *drinking together*.

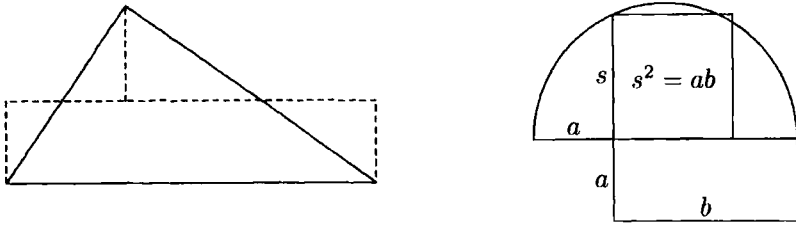


FIGURE 1. Left: turning a triangle into a rectangle. Right: turning a rectangle into a square ($s^2 = ab$).

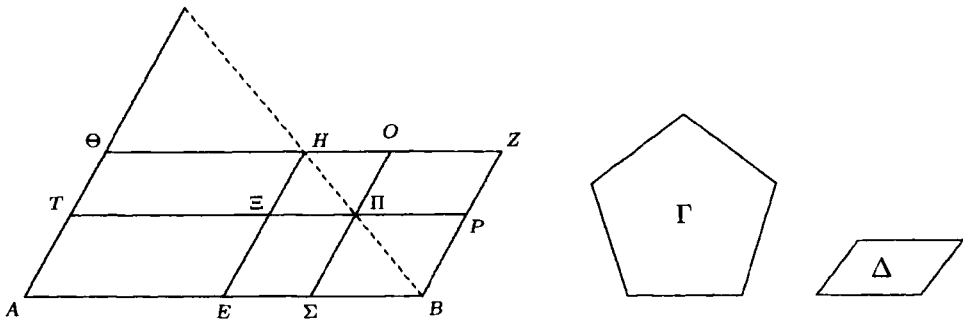


FIGURE 2. Application with defect. Euclid, Book 6, Proposition 28.

problem does not have a solution for all given lines and areas, since the largest parallelogram that can be formed under these conditions is the one whose base is half of the given line (Book 6, Proposition 26). Assuming that condition, let AB be the given line, Γ the given polygonal region, and Δ the given parallelogram shape. The dashed line from B makes the same angle with AB that the diagonal of the parallelogram Δ makes with its base. The line $A\Theta$ is drawn to make the same angle as the corresponding sides of Δ . Then any parallelogram having its sides along AB and $A\Theta$ and opposite corner on the dashed line will automatically generate a “defect” that is similar to Δ . The remaining problem is to find the one that has the same area as Γ . That is achieved by constructing the parallelogram $H\Xi\Pi O$ similar to Δ and equal to the difference between $AEH\Theta$ and Γ .

The Greek word for application is *parabolê*. Proclus cites Eudemus in asserting that the solution of the application problems was an ancient discovery of the Pythagoreans, and that they gave them the names *ellipse* and *hyperbola*, names that were later transferred to the conic curves by Apollonius. This version of events is also confirmed by Pappus. We shall see the reason for the transfer below.

Although most of Euclid’s theorems have obvious interest from the point of view of anyone curious about the world, the application problems raise a small mystery. Why were the Pythagoreans interested in them? Were they merely a refinement of the transformation problems? Why would anyone be interested in applying an area so as to have a defect or excess of a certain shape? Without restriction on the shape of the defect or excess, the application problem does not

have a unique solution. Were the additional conditions imposed simply to make the problem determinate? Some historians have speculated that there was a further motive.

In the particular case when the excess or defect is a square, these problems amount to finding two unknown lengths given their sum and product (application with deficiency) or given their difference and product (application with excess). In modern terms these two problems amount to quadratic equations. Some authors have argued that this “geometric algebra” was a natural response to the discovery of incommensurable magnitudes, described in Chapter 7, indeed a logically necessary response. On this point, however, many others disagree. Gray, for example, says that, while the discovery of incommensurables did point out a contradiction in a naive approach to ratios, “it did not provoke a foundational crisis.” Nor did it force the Pythagoreans to recast algebra as geometry:

There is no logical necessity about it. It would be quite possible to persevere with an arithmetic of natural numbers to which was adjoined such new quantities as, say, arose in the solution of equations. There is nothing more intelligible about a geometric segment than a root of an equation, unless you have already acquired a geometric habit of thought. Rather than turning from algebra to geometry, I suggest that the Greeks were already committed to geometry. [Gray, 1989, p. 16]

1.4. Challenges to Pythagoreanism: unsolved problems. Supposing that this much was known to the early Pythagoreans, we can easily guess what problems they would have been trying to solve. Having learned how to convert any polygon to a square of equal area, they would naturally want to do the same with circles and sectors and segments of circles. This problem was known as *quadrature (squaring) of the circle*. Also, having solved the transformation problems for a plane, they would want to solve the analogous problems for solid figures, in other words, to convert a polyhedron to a cube of equal volume. Finding the cube would be interpreted as finding the length of its side. Now, the secret of solving the planar problem was to triangulate a polygon, construct a square equal to each triangle, then add the squares to get bigger squares using the Pythagorean theorem. By analogy, the three-dimensional program would be to cut a polyhedron into tetrahedra, convert any tetrahedron into a cube of equal volume, then find a way of adding cubes analogous to the Pythagorean theorem for adding squares. The natural first step of this program (as we imagine it to have been) was to construct a cube equal to the double of a given cube, the problem of *doubling the cube*, just as we imagined that doubling a square may have led to the Pythagorean theorem.

Yet another example of such a problem is that of dividing an arc (or angle) into equal parts. If we suppose that the Pythagoreans knew how to bisect arcs (Proposition 9 of Book 1 of the *Elements*) and how to divide a line into any number of equal parts (Proposition 9 of Book 6), this asymmetry between their two basic figures—lines and circles—would very likely have been regarded as a challenge. The first step in this problem would have been to divide an arc into three equal parts, the problem of *trisection of the angle*.

The three problems just listed were mentioned by later commentators as an important challenge to all geometers. To solve them, geometers had to enlarge their set of basic objects beyond lines and planes. They were rather conservative in

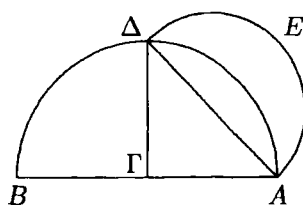


FIGURE 3. Hippocrates' quadrature of a lune, according to Simplicius.

doing so, first invoking familiar surfaces such as cones and cylinders, which could be generated by moving lines on circles, and intersecting them with planes so as to get the conic sections that we know as the ellipse, parabola, and hyperbola. These curves made it possible to solve two of the three problems (trisecting the angle and doubling the cube). Later, a number of more sophisticated curves were invented, among them spirals, the cissoid, and the quadratrix. This last curve got its name from its use in squaring the circle. Although it is not certain that the Pythagoreans had a program like the one described here, it is known that all three of these problems were worked on in antiquity.

Squaring the circle. Proclus mentions Hippocrates of Chios as having discovered the quadratures of lunes. In fact this mathematician (ca. 470–ca. 410 BCE), who lived in Athens at the time of the Peloponnesian War (430–404), is said to have worked on all three of the classical problems. A lune is a figure resembling a crescent moon: the region inside one of two intersecting circles and outside the other. In the ninth volume of his commentary on Aristotle's books on physics, the sixth-century commentator Simplicius discusses several lunes that Hippocrates squared, including the one we are about to discuss. After detailing the criticism by Eudemus of earlier attempts by Antiphon (480–411) to square the circle by the kind of polygonal approximation we discussed in Chapter 9, Simplicius reports one of Hippocrates' quadratures (Fig. 3), based on Book 12 of Euclid's *Elements*. The result needed is that semicircles are proportional to the squares on their diameters.

Simplicius' reference to Book 12 of Euclid is anachronistic, since Hippocrates lived before Euclid; but it was probably well known that similar segments are proportional to the squares on their bases. Even that theorem is not needed here, except in the case of semicircles, and that special case is easy to derive from the theorem for whole circles. The method of Hippocrates does not achieve the quadrature of a whole circle; we can see that his procedure works because the "irrationalities" of the two circles cancel each other when the segment of the larger circle is removed from the smaller semicircle.

In his essay *On Exile*, Plutarch reports that the philosopher Anaxagoras worked on the quadrature of the circle while imprisoned in Athens. (He was brought there by Pericles, who was eventually compelled to send him away.) Other attempts are reported, one by Dinostratus (ca. 390–ca. 320 BCE), who is said to have used the curve called (later, no doubt) the *quadratrix* (*squarer*), invented by Hippias of Elis (ca. 460–ca. 410 BCE) for the purpose of trisecting the angle. It is discussed below in that connection.

Doubling the cube. Although the problem of doubling the cube fits very naturally into what we have imagined as the Pythagorean program, some ancient authors

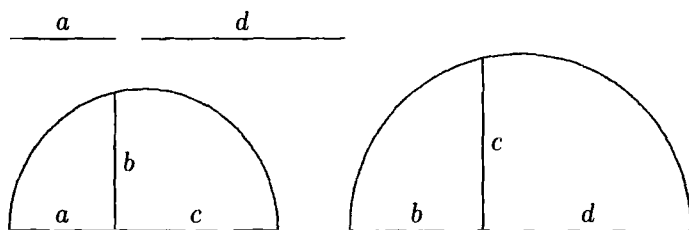


FIGURE 4. The problem of two mean proportionals: Given a and d , find b and c .

gave it a more exotic origin. In *The Utility of Mathematics*, the commentator Theon of Smyrna, who lived around the year 100 CE, discusses a work called *Platonicus* that he ascribes to Eratosthenes, to the effect that the citizens of Delos (the island that was the headquarters of the Athenian Empire) consulted an oracle in order to be relieved of a plague, and the oracle told them to double the size of an altar (probably to Apollo). Plagues were common in ancient Greece; one is described in Sophocles' *Oedipus the King*, and another decimated Athens early in the Peloponnesian War, claiming Pericles as one of its victims. According to Theon, Eratosthenes depicted the Delians as having turned for technical advice to Plato, who told them that the altar was not the point: The gods really wanted the Delians to learn geometry better. In his commentary on Archimedes' work on the sphere and cylinder, Eutocius gives another story, also citing Eratosthenes, but he says that Eratosthenes told King Ptolemy in a letter that the problem arose on the island of Crete when King Minos ordered that a tomb built for his son be doubled in size.

Whatever the origin of the problem, both Proclus and Eutocius agree that Hippocrates was the first to reduce it to the problem of two mean proportionals. The Pythagoreans knew that the mean proportional between any two square integers is an integer, for example, $\sqrt{16 \cdot 49} = 28$ and that between any two cubes such as 8 and 216 there are two mean proportionals (Euclid, Book 8, Propositions 11 and 12); for example, $8 : 24 :: 24 : 72 :: 72 : 216$. If two mean proportionals could be found between two cubes—as seems possible, since every volume can be regarded as the cube on some line—the problem would be solved. It would therefore be natural for Hippocrates to think along these lines when comparing two cubes. Eutocius, however, was somewhat scornful of this reduction, saying that the new problem was just as difficult as the original one. That claim, however, is not true: One can easily draw a figure containing two lines and their mean proportional (Fig. 1): the two parts of the diameter on opposite sides of the endpoint of the half-chord of a circle and the half-chord itself. The only problem is to get two such figures with the half-chord and one part of the diameter reversing roles between the two figures and the other parts of the diameters equal to the two given lines, as shown in Fig. 4. It is natural to think of using two semicircles for this purpose and moving the chords to meet these conditions.

In his commentary on the treatise of Archimedes on the sphere and cylinder, Eutocius gives a number of solutions to this problem, ascribed to various authors, including Plato. The earliest one that he reports is due to Archytas (ca. 428–350 BCE). This solution requires intersecting a cylinder with a torus and a cone.



FIGURE 5. The three conic sections, according to Menaechmus.

The three surfaces intersect in a point from which the two mean proportionals can easily be determined. A later solution by Menaechmus may have arisen as a simplification of Archytas' rather complicated construction. It requires intersecting two cones, each having a generator parallel to a generator of the other, with a plane perpendicular to both generators. These intersections form two conic sections, a parabola and a rectangular hyperbola; where they intersect, they produce the two mean proportionals.

If Eutocius is correct, the conic sections first appeared, but not with the names they now bear, in the late fourth century BCE. Menaechmus created these sections by cutting a cone with a plane perpendicular to one of its generators. When that is done, the shape of the section depends on the vertex angle of the cone. If that angle is acute, the section will be an ellipse; if it is a right angle, the section will be a parabola; if it is obtuse, the section will be a hyperbola. In his commentary on Archimedes' treatise on the sphere and cylinder, Eutocius tells how he happened to find a work written in the Doric dialect which seemed to be a work of Archimedes. He mentions in particular that instead of the word *parabola*, used since the time of Apollonius, the author used the phrase *section of a right-angled cone*, and instead of *hyperbola*, the phrase *section of an obtuse-angled cone*. Since Proclus refers to "the conic section triads of Menaechmus," it is inferred that the original names of the conic sections were *oxytomē* (sharp cut), *orthotomē* (right cut), and *amblytomē* (blunt cut), as shown in Fig. 5. However, Menaechmus undoubtedly thought of the cone as the portion of the figure from the vertex to some particular circular base. In particular, he wouldn't have thought of the hyperbola as having two nappes, as we now do.

How Apollonius came to give them their modern names a century later is described below. Right now we shall look at the consequences of Menaechmus' approach and see how it enabled him to solve the problem of two mean proportionals. It is very difficult for a modern mathematician to describe this work without breaking into modern algebraic notation, essentially using analytic geometry. It is very natural to do so; for Menaechmus, if Eutocius reports correctly,⁷ comes very close to stating his theorem in algebraic language.

⁷ That is a big "if." Eutocius clearly had read Apollonius; Menaechmus, just as clearly, could not have done so.

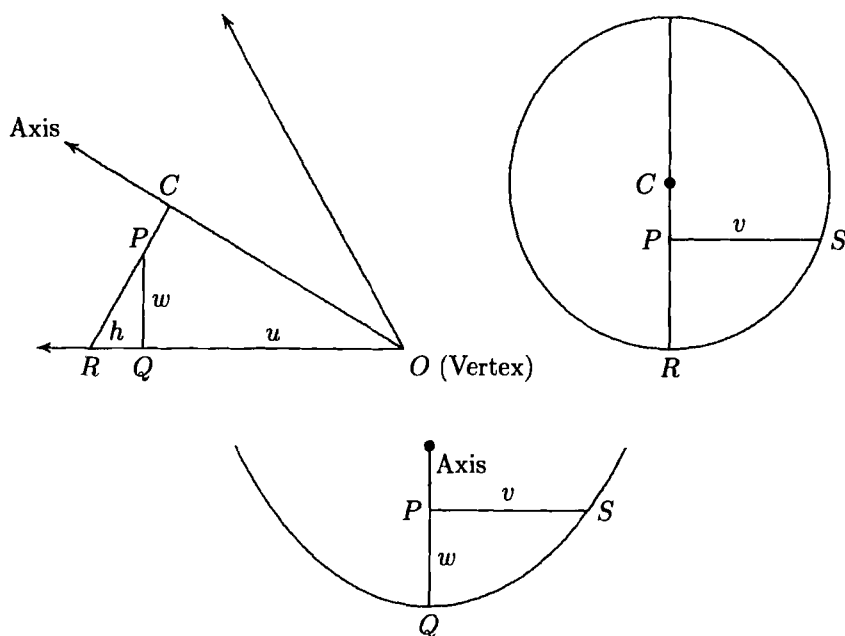


FIGURE 6. Sections of a cone. Top left: through the axis. Top right: perpendicular to the axis. Bottom: perpendicular to the generator OR at a point Q lying at distance u from the vertex O . The fundamental relation is $v^2 = h^2 + 2uh - w^2$. The length h has a fixed ratio to w , depending only on the shape of the triangle OCR .

We begin by looking at a general conic section, shown in Fig. 6. When a cone is cut by a plane through its axis, the resulting figure is simply a triangle. The end that we have left open by indicating with arrows the direction of the axis and two generators in this plane would probably have been closed off by Menaechmus. If it is cut by a plane perpendicular to the axis, the result is a circle. The conic section itself is obtained as the intersection with a plane perpendicular to one of its generators at a given distance (marked u in the figure) from the vertex. The important relation needed is that between the length of a horizontal chord (double the length marked v) in the conic section and its height (marked w) above the generator that has been cut. Using only similar triangles and the fact that a half chord in a circle is the mean proportional between the segments of the diameter through its endpoint, Menaechmus would easily have derived the fundamental relation

$$(1) \quad v^2 = h^2 + 2uh - w^2.$$

Although we have written this relation as an equation with letters in it, Menaechmus would have been able to describe what it says in terms of the lines v , u , h , and w , and squares and rectangles on them. He would have known the value of the ratio h/w , which is determined by the shape of the triangle ROC . In our terms $h = w \tan(\varphi/2)$, where φ is the vertex angle of the cone.

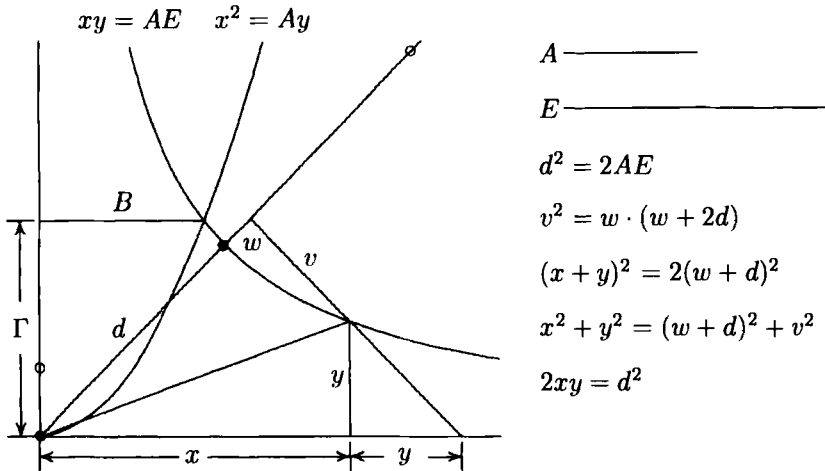


FIGURE 7. One of Menaechmus' solutions to the problem of two mean proportionals, as reported by Eutocius.

The simplest case is that of the parabola, where the vertex angle is 90° and $h = w$. In that case the relation between v and w is

$$v^2 = 2uw.$$

In the problem of putting two mean proportionals B and Γ between two lines A and E , Menaechmus took this u to be $\frac{1}{2}A$, so that $v^2 = Aw$.

The hyperbola Menaechmus needed for this problem was a rectangular hyperbola, which results when the triangle ROC is chosen so that $\overline{RC}^2 = 2\overline{OC}^2$, and therefore $\overline{OR}^2 = 3\overline{OC}^2$. Such a triangle is easily constructed by extending one side of a square to the same length as the diagonal and joining the endpoint to the opposite corner of the square. In any triangle of this shape the legs are the side and diagonal of a square. For that case Menaechmus would have been able to show that the relation

$$v^2 = w(w + 2d)$$

holds, where d is the diagonal of a square whose side is u . To solve the problem of two mean proportionals, Menaechmus took $u = \sqrt{AE}$, that is, the mean proportional between A and E . Menaechmus' solution is shown in Fig. 7.

The solution fits perfectly within the framework of Pythagorean-Euclidean geometry, yet people were not satisfied with it. The objection to it was that the data and the resulting figure all lie within a plane, but the construction requires the use of cones, which cannot be contained in the plane.

Trisecting the angle. The practicality of trisecting an angle is immediately evident: It is the vital first step on the way to dividing a circular arc into any number of equal pieces. If a right angle can be divided into n equal pieces, a circle also can be divided into n equal pieces, and hence the regular n -gon can be constructed. The success of the Pythagoreans in constructing the regular pentagon must have encouraged them to pursue this program. It is possible to construct the regular n -gon for $n = 3, 4, 5, 6, 8, 10$, but not 7 or 9. The number 7 is awkward, being the only prime between 5 and 10, and one could expect to have difficulty constructing

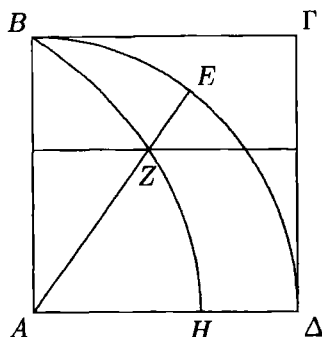


FIGURE 8. The quadratrix of Hippias.

the regular heptagon. Surprisingly, however, the regular 17-sided polygon *can* be constructed using only compass and straightedge. Since $9 = 3 \cdot 3$, it would seem natural to begin by trying to construct this figure, that is, to construct an angle of 40° . That would be equivalent to constructing an angle of 20° , hence trisecting the angles of an equilateral triangle.

Despite the seeming importance of this problem, less has been written about the ancient attempts to solve it than about the other two problems. For most of the history we are indebted to two authors. In his commentary on Euclid's *Elements*, Proclus mentions the problem and says that it was solved by Nicomedes using his conchoid and by others using the quadratrices of Hippias and Nicomedes. In Book 4 of his *Synagōgē* (*Collection*), Pappus says that the circle was squared using the curve of Dinostratus and Nicomedes. He then proceeds to describe that curve, which is the one now referred to as the quadratrix of Hippias.⁸

The quadratrix is described physically as follows. The radius of a circle rotates at a uniform rate from the vertical position AB in Fig. 8 to the horizontal position $A\Delta$, while in exactly the same time a horizontal line moves downward at a constant speed from the position $B\Gamma$ to the position $A\Delta$. The point of intersection Z traces the curve BZH , which is the quadratrix. The diameter of the circle is the mean proportional between its circumference and the line AH . Unfortunately, H is the one point on the quadratrix that is not determined, since the two intersecting lines coincide when they both reach $A\Delta$. This point was noted by Pappus, citing an earlier author named Sporos. In order to draw the curve, which is mechanical, you first have to know the ratio of the circumference of a circle to its diameter. But if you knew that, you would already be able to square the circle. One can easily see, however, that since the angle $Z\Delta\Delta$ is proportional to the height of Z , this curve makes it possible to divide an angle into any number of equal parts.

Pappus also attributed a trisection to Menelaus of Alexandria. Pappus gave a classification of geometric construction problems in terms of three categories: planar, solid, and [curvi]linear. The first category consisted of constructions that used only straight lines and circles, the second those that used conic sections. The last,

⁸ Hippias should be thankful for Proclus, without whom he would apparently be completely forgotten, as none of the other standard commentators discuss him, except for a mention in passing by Diogenes Laertius in his discussion of Thales. Allman (1889, pp. 94–95) argues that the Hippias mentioned in connection with the quadratrix is not the Hippias of Elis mentioned in the Eudemian summary, and other historians have agreed with him, but most do not.

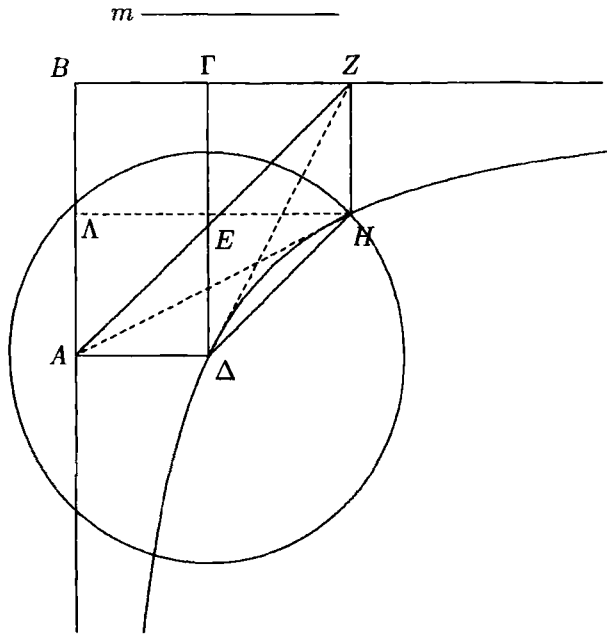


FIGURE 9. Pappus' construction of a *neûsis* using a rectangular hyperbola.

catch-all category consisted of problems requiring all manner of more elaborate and less regular curves, which were harder to visualize than the first two and presumably required some mechanical device to draw them. Pappus says that some of these curves come from locus problems, and lists the inventors of some others, among them a curve that Menelaus called the *paradox*. Other spirals of the same type, he says, are the quadratrices, the conchoids, and the cissoids. He goes on to say that geometers regard it as a major defect when a planar problem is solved using conics and other curves.

Based on this classification of problems, the first geometers were unable to solve the abovementioned problem of [trisecting] the angle, which is by nature a solid problem, through planar methods. For they were not yet familiar with the conic sections; and for that reason they were at a loss. But later they trisected the angle through conics, using the convergence described below.

The word *convergence* (*neûsis*) comes from the verb *neûein*, one of whose meanings is *to incline toward*. In this particular case, it refers to the following construction. We are given a rectangle $AB\Gamma\Delta$ and a prescribed length m . It is required to find a point E on $\Gamma\Delta$ such that when AE is drawn and extended to meet the extension of $B\Gamma$ at a point Z , the line EZ will have length m . The construction is shown in Fig. 9, where the circle drawn has radius m and the hyperbola is rectangular, so that $A\Delta \cdot \Gamma\Delta = \Lambda H \cdot ZH$.

Given the *neûsis*, it becomes a simple matter to trisect an angle, as Pappus pointed out. Given any acute angle, label its vertex A , choose an arbitrary point Γ on one of its sides, and let Δ be the foot of the perpendicular from Γ to the other side of the angle. Complete the rectangle $AB\Gamma\Delta$, and carry out the *neûsis* with $m = 2A\Gamma$. Then let H be the midpoint of ZE , and join ΓH , as shown in Fig. 10.

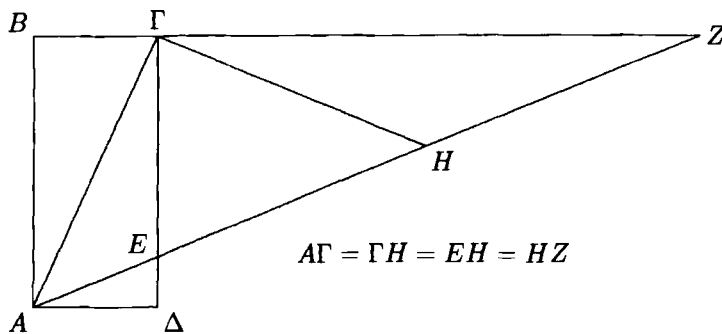


FIGURE 10. Trisection of an arbitrary angle by *neûsis* construction.

A mechanical (curvilinear) solution of the neûsis problem. Finding the point E in the *neûsis* problem is equivalent to finding the point Z . Either point allows the line $A EZ$ to be drawn. Now one line that each of these points lies on is known. If some other curve that Z must lie on could be drawn, the intersection of that curve with the line $B\Gamma$ would determine Z and hence solve the *neûsis* problem. If we use the condition that the line ZE must be of constant length, we have a locus-type condition for Z , and it is easy to build a device that will actually draw this locus. What is needed is the T-shaped frame shown in Fig. 11, consisting of two pieces of wood or other material meeting at right angles. The horizontal part of the T has a groove along which a peg (shown as a hollow circle in the figure) can slide. The vertical piece has a fixed peg (shown as a solid circle) at distance $A\Delta$ from its top. Onto this frame a third piece is fitted with a fixed peg (the hollow circle) at distance m from its end and a groove between the peg and the other end that fits over the peg on the vertical bar. The frame is then laid down with its horizontal groove over the line $\Gamma\Delta$ and its fixed peg over A . When the moving piece is fitted over the frame so that its peg slides along the horizontal groove over $\Gamma\Delta$ and its groove slides over the peg at A , its endpoint (where a stylus is located to draw the curve) traces the locus on which Z must lie. The point Z lies where that locus meets the extension of $B\Gamma$. In practical terms, such a device can be built, but the rigid pegs must be located at exactly the distance from the ends determined by the rectangle and the fixed distance given in the *neûsis* problem. Thus the device must be modified by moving the pegs to the correct locations for each particular problem. If oxymoron is permitted, we might say that the practical value of this device is mostly theoretical. The locus it draws is the *conchoid of Nicomedes*, mentioned by Pappus and Proclus.

Because of the objections reported by Pappus to the use of methods that were more elaborate than the problems they were intended to solve, the search for planar (ruler-and-compass) solutions to these problems continued for many centuries. It was not until the 1830s that it was proved that no ruler-and-compass solution exists. The proof had no effect on the cranks of the world, of course. The problems continue to be of interest since that time, and not only to cranks who imagine they have solved them. Felix Klein, a leading German mathematician and educator in

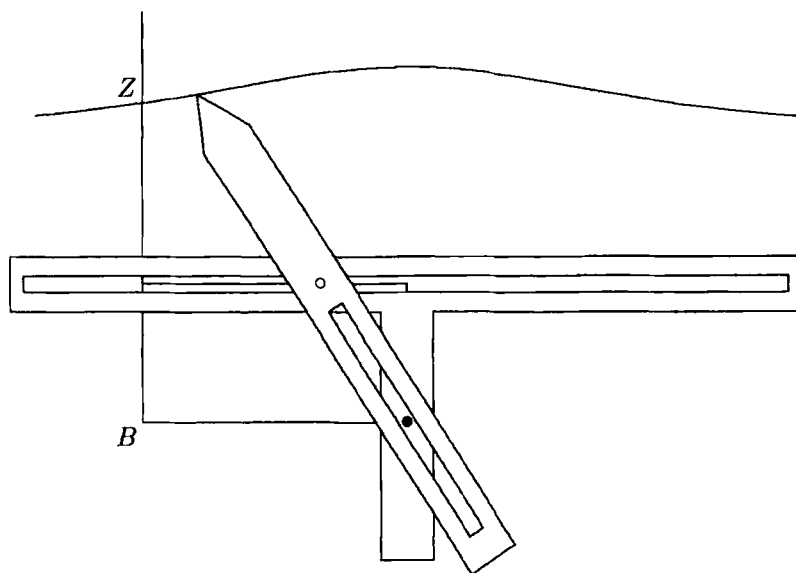


FIGURE 11. A mechanical device for drawing the conchoid of Nicomedes.

the late nineteenth and early twentieth centuries, urged that they be studied as a regular part of the curriculum (Beman and Smith, 1930).

1.5. Challenges to Pythagoreanism: the paradoxes of Zeno of Elea. Although we have some idea of the geometric results proved by the Pythagoreans, our knowledge of their interpretation of these results is murkier. How did they conceive of geometric entities such as points, lines, planes, and solids? Were these objects physically real or merely ideas? What properties did they have? Some light is shed on this question by the philosophical critics of Pythagoreanism, one of whom has become famous for the paradoxes he was able to spin out of Pythagorean principles.

In the Pythagorean philosophy the universe was generated by numbers and motion. That these concepts needed to be sharpened up became clear from critics of the Pythagorean school. It turned out that the Pythagorean view of geometry and number contained paradoxes within itself, which were starkly pointed out by the philosopher Zeno of Elea. Zeno died around 430 BCE, and we do not have any of his works to rely on, only expositions of them by other writers. Aristotle, in particular, says that Zeno gave four puzzles about motion, which he called the Dichotomy (division), the Achilles, the Arrow, and the Stadium. Here is a summary of these arguments in modern language, based on Book 6 of Aristotle's *Physics*.

The Dichotomy. Motion is impossible because before an object can arrive at its destination it must first arrive at the middle of its route. But before it can arrive at the middle, it must travel one-fourth of the way, and so forth. Thus we see that the object must do infinitely many things in a finite time in order to move.

The Achilles. (This paradox is apparently so named because in Homer's *Iliad* the legendary warrior Achilles chased the Trojan hero Hector around the walls of Troy, overtook him, and killed him.) If given a head start, the slower runner will never be overtaken by the faster runner. Before the two runners can be at the same point at the same instant, the faster runner must first reach the point from which the slower runner started. But at that instant the slower runner will have reached another

point ahead of the faster. Hence the race can be thought of as beginning again at that instant, with the slower runner still having a head start. The race will “begin again” in this sense infinitely many times, with the slower runner always having a head start. Thus, as in the dichotomy, infinitely many things must be accomplished in a finite time in order for the faster runner to overtake the slower.

The Arrow. An arrow in flight is at rest at each instant of time. That is, it does not move from one place to another during that instant. But then it follows that it cannot traverse any positive distance because successive additions of zero will never result in anything but zero.

The Stadium. (In athletic stadiums in Greece the athletes ran from the goal, around a halfway post and then back. This paradox seems to have been inspired by imagining two lines of athletes running in opposite directions and meeting each other.) Consider two parallel line segments of equal length moving toward each other with equal speeds. The speed of each line is measured by the number of points of space it passes over in a given time. But each point of one line passes *twice* that many points of the other line in the same time as the two lines move past each other. Hence the velocity of the line must equal its double, which is absurd.

The Pythagoreans had built their system on lines “made of” points, and now Zeno was showing them that space cannot be “made of” points in the same way that a building can be made of bricks. For assuredly the number of points in a line segment cannot be finite. If it were, the line would not be infinitely divisible as the dichotomy and Achilles paradoxes showed that it must be; moreover, the stadium paradox would show that the number of points in a line segment equals its double. There must therefore be an infinity of points in a line. But then each of these points must take up no space; for if each point occupied some space, an infinite number of them would occupy an infinite amount of space. But if points occupy no space, how can the arrow, whose tip is at a single point at each instant of time, move through a *positive* quantity of space? A continuum whose elements are points was needed for geometry, yet it could not be thought of as being made up of points in the way that discrete collections are made up of individuals.

1.6. Challenges to Pythagoreanism: incommensurables. The difficulties pointed out by Zeno affected the intuitive side of geometry. The challenge they posed may have been an impetus to the kind of logical rigor that we know as Euclidean. There is, however, an even stronger impetus to that rigor, one that was generated from within Pythagorean geometry. To the modern mathematician, this second challenge to Pythagorean principles is much more relevant and interesting than the paradoxes of Zeno. That challenge is the problem of incommensurables, which led ultimately to the concept of a real number.

The existence of incommensurables throws doubt on certain oversimplified proofs of geometric proportion. When two lines or areas are commensurable, one can describe their ratio as, say, $5 : 7$, meaning that there is a common measure such that the first object is five times this measure and the second is seven times it. A proportion such as $a : b :: c : d$, then, is the statement that ratios $a : b$ and $c : d$ are both represented by the same pair of numbers.

This theory of proportion is extremely important in geometry if we are to have such theorems as Proposition 1 of Book 6 of Euclid’s *Elements*, which says that the areas of two triangles or two parallelograms having the same height are proportional

to their bases, or the theorem (Book 12, Proposition 2) that the areas of two circles are proportional to the squares on their diameters. Even the simplest constructions, such as the construction of a square equal in area to a given rectangle or the application problems mentioned above, may require the concept of proportionality of lines. Because of the importance of the theory of proportion for geometry, the discovery of incommensurables made it imperative to give a definition of proportion without relying on a common measure to define a ratio.

Fowler (1998) argues for the existence of a Pythagorean theory of proportion based on *anthypharesis*, the mutual subtraction procedure we have now described many times.⁹ He makes a very telling point (p. 18) in citing a passage from Aristotle's *Topics* where the assertion is made that having the same *antanaíresis* is tantamount to having the same ratio. Fowler takes *antanaíresis* to be a synonym of *anthypharesis*. Like Gray (quoted above), Knorr (1975) argues that the discovery of irrationals was not a major "scandal," and that it was not responsible for the "geometric algebra" in Book 2 of Euclid. While arguing that incommensurability forced some modifications in the way the Pythagoreans thought about physical magnitudes, he says (p. 41):

It is thus thoroughly obvious that, far from being in a state of paralysis, fifth- and fourth-century geometers proceeded with their studies of similar figures as if they were still unaware of the foundational consequences of the existence of incommensurable lines.

1.7. The influence of Plato. Plato is still held in high esteem by philosophers, and it is well recognized that his philosophy contains a strong mathematical element. But since Plato was a follower of Socrates, who was almost entirely concerned with questions of ethics and the right conduct of life,¹⁰ his interest in mathematical questions needs to be explained. Born in 427 BCE, Plato served in the Athenian army during the Peloponnesian War. He was also a devoted follower of Socrates. Socrates enjoyed disputation so much and was so adept at showing up the weakness in other people's arguments that he made himself very unpopular. When Athens was defeated in 404 BCE, Plato sided with the party of oligarchs who ruled the city temporarily. When the democratic rule was restored, the citizens took revenge on their enemies, among whom they counted Socrates. Plato was devastated by the trial and execution of Socrates in 399 BCE. He left Athens and traveled to Italy, where he became acquainted with the Pythagorean philosophy. He seems to have met the Pythagorean Philolaus in Sicily in 390. He also met the Pythagorean Archytas at Tarentum (where some Pythagoreans had fled to escape danger at Croton). Plato returned to Athens and founded the Academy in 387 BCE. There he hoped to train the young men¹¹ for public service and establish good government. At the behest of Archytas and a Syracusan politician named Dion, brother-in-law of the ruler Dionysus I, Plato made several trips to Syracuse, in Sicily, between 367 and 361 BCE, to act as advisor to Dionysus II. However, there was virtual civil war between Dion and Dionysus, and Plato was arrested and nearly executed. Diogenes

⁹ Fowler avoids as far as possible using the phrase *Euclidean algorithm*.

¹⁰ The Socrates depicted by Plato is partly a literary device through which Plato articulated his own thoughts on many subjects that the historical Socrates probably took little notice of.

¹¹ In his writing, especially *The Republic*, Plato argues for equal participation by women in government. There is no record of any female student at his Academy, however. His principles were far in advance of what the Athenians would tolerate in practice.

Laertius quotes a letter allegedly from Archytas to Dionysus urging that Plato be released. Plato returned to the Academy in 360 and remained there for the last 13 years of his life. He died in 347.

Archytas. Archytas, although a contemporary of Plato, is counted paradoxically among the “pre-Socratics” in philosophy; but that is because he worked outside Athens and continued the earlier Pythagorean tradition. Archytas’ solution of the problem of two mean proportionals using two half-cylinders intersecting at right angles was mentioned above. In his *Symposium Discourses*, Plutarch claimed that

for that reason Plato also lamented that the disciples of Eudoxus, Archytas, and Menaechmus attacked the duplication of a solid by building tools and machinery hoping to get two ratios through the irrational, by which it might be possible to succeed, [saying that by doing so they] immediately ruined and destroyed the good of geometry by turning it back toward the physical and not directing it upward or striving for the eternal and incorporeal images, in which the god is ever a god.

Although the sentiment Plutarch ascribes to Plato is consistent with the ideals expressed in the *Republic*, Eutocius reports one such mechanical construction as being due to Plato himself. From his upbringing as a member of the Athenian elite and from the influence of Socrates, Plato had a strong practical streak, concerned with life as it is actually lived.¹² Platonic idealism in the purely philosophical sense does not involve idealism in the sense of unrealistic striving for perfection.

Archytas and Philolaus provided the connection between Pythagoras and Plato, whose interest in mathematics began some time after the death of Socrates and continued for the rest of his life. Mathematics played an important role in the curriculum of his Academy and in the research conducted there, and Plato himself played a leading role in directing that research. Lasserre (1964, p. 17) believes that the most important mathematical work at the Academy began with the arrival of Theaetetus in Athens around 375 and ended with Eudoxus’ departure for Cnidus around 350.

The principle that knowledge can involve only eternal, unchanging entities led Plato to some statements that sound paradoxical. For example, in Book 7 of the *Republic* he writes:

Thus we must make use of techniques such as geometry when we take up astronomy and let go of the things in the heavens if we really intend to create something intrinsically useful and practical in the soul by mastering astronomy.

If Plato’s mathematical concerns seem to be largely geometrical, that is probably because he encountered Pythagoreanism at the time when the challenges discussed above were still current topics. (Recall the quotation from the *Republic* in

¹² In the famous allegory of the cave in Book 7 of the *Republic*, Plato depicts the nonphilosophical person as living in a cave with feet in chains, seeing only flickering shadows on the wall of the cave, while the philosopher is the person who has stepped out of the cave into the bright sunshine and wishes to communicate that reality to the people back in the cave. While he encouraged his followers to “think outside the cave,” his trips to Syracuse show that he understood the need to make philosophy work inside the cave, where everyday life was going on.

Chapter 3, where he laments the lack of public support for research into solid geometry.) There is a long-standing legend that Plato's Academy bore the following sign above its entrance:¹³

ΑΓΕΩΜΕΤΡΗΤΟΣ ΜΗΔΕΙΣ ΕΙΣΙΤΩ

("Let no ungeometrical person enter.") If Plato was indeed more concerned with geometry than with arithmetic, there is an obvious explanation for his preference: The imperfections of the real world relate entirely to geometry, not at all to arithmetic. For example, it is sometimes asserted that there are no examples of exact equality in the real world. But in fact, there are many. Those who make the assertion always have in mind continuous magnitudes, such as lengths or weights, in other words, geometrical concepts. Where arithmetic is concerned, exact equality is easy to achieve. If I have \$11,328.75 in the bank, and my neighbor has \$11,328.75 in the bank, the two of us have *exactly* the same amount of money. Our bank accounts are interchangeable for all monetary purposes. But Plato's love for geometry should not be overemphasized. In his ideal curriculum, described in the *Republic*, arithmetic is still regarded as the primary subject.

1.8. Eudoxan geometry. To see why the discovery of incommensurables created a problem for the Pythagoreans, consider the following conjectured early proof of a fundamental result in the theory of proportion: the proposition that two triangles having equal altitudes have areas proportional to their bases. This assertion is half of Proposition 1 of Book 6 of Euclid's *Elements*. Let ABC and ACD in Fig. 12 be two triangles having the same altitude. Euclid draws them as having a common side, but that is only for convenience. This positioning causes no loss in generality because of the proposition that any two triangles of equal altitude and equal base have equal areas, proved as Proposition 38 of Book 1.

Suppose that the ratio of the bases $BC : CD$ is $2 : 3$, that is, $3BC = 2CD$. Extend BD leftward to H so that $BC = BG = GH$, producing triangle AHC , which is three times triangle ABC . Then extend CD rightward to K so that $CD = DK$, yielding triangle ACK equal to two times triangle ACD . But then, since $GC = 3BC = 2CD = CK$, triangles AGC and ACK are equal. Since $AGC = 3ABC$ and $ACK = 2ACD$, it follows that $ABC : ACD = 2 : 3$. We, like Euclid, have no way of actually *drawing* an unspecified number of copies of a line, and so we are forced to *illustrate* the argument using specific numbers (2 and 3 in the present case), while expecting the reader to understand that the argument is completely general.

An alternative proof could be achieved by finding a common measure of BC and CD , namely $\frac{1}{2}BC = \frac{1}{3}CD$. Then, dividing the two bases into parts of this length, one would have divided ABC into two triangles, ACD into three triangles, and all five of the smaller triangles would be equal. But both of these arguments fail if no integers m and n can be found such that $mBC = nCD$, or (equivalently) no common measure of BC and CD exists. This proof needs to be shored up, but how is that to be done?

¹³ These words are the earliest version of the legend, which Fowler (1998, pp. 200–201) found could not be traced back earlier than a scholium attributed to the fourth-century orator Sopatros. The commonest source cited for this legend is the twelfth-century Byzantine Johannes Tzetzes, in whose *Chiliades*, VIII, 975, one finds Μηδεις ἀγεωμέτρητος εἰσὶτω μου τὴν στέγην. "Let no ungeometrical person enter my house."

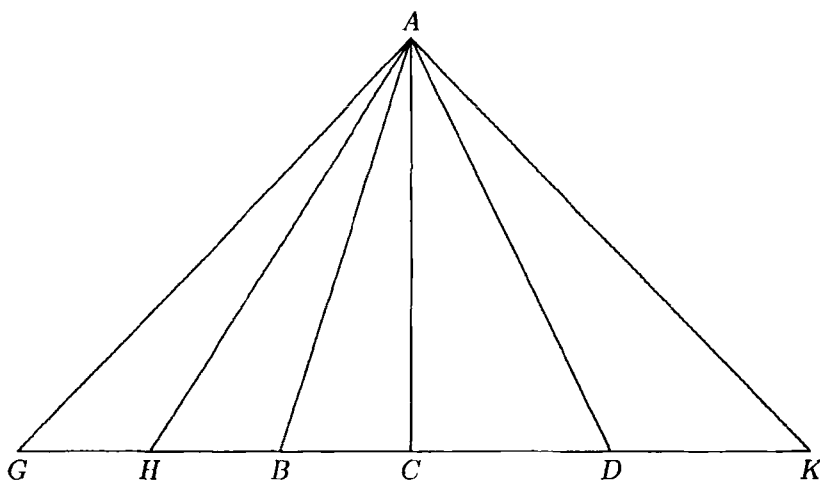


FIGURE 12. A fundamental theorem in the theory of proportion.
Proposition 1 of Book 6 of the *Elements*.

The solution to the difficulty was provided by Eudoxus of Cnidus (ca. 407–354 BCE), whom Diogenes Laërtius describes as “astronomer, geometer, physician, and lawgiver.” He learned geometry from Archytas. Diogenes Laërtius cites another commentator, named Sotion, who said that Eudoxus spent two months in Athens and attended lectures by Plato. Because of his poverty, he could not afford to live in Athens proper. He lived at the waterfront, known as the Piraeus, supported by a physician named Theomedus, and walked 11 km from there into Athens. Then, with a subsidy from friends, he went to Egypt and other places, and finally returned, “crammed full of knowledge,” to Athens, “some say, just to annoy Plato for snubbing him earlier.” Plato was not in Eudoxus’ league as a mathematician; and if Eudoxus felt that Plato had patronized him in his earlier visit, perhaps because Plato and his other students were wealthy and Eudoxus was poor, his desire to return and get his own back from Plato is quite understandable. He must have made an impression on Plato on his second visit. In his essay *On Socrates’ Daemon*, Plutarch reports that when the Delians consulted Plato about doubling the cube, in addition to advising them to study geometry, he told them that the problem had already been solved by Eudoxus of Cnidus and Helicon of Cyzicus. If true, this story suggests that the Delians appealed to Plato after Eudoxus had left for Cnidus, around 350. By that time Plato was a very old man, and perhaps mellowed than he had been a quarter-century earlier during Eudoxus’ first stay in Athens. In Cnidus, Eudoxus made many astronomical observations that were cited by the astronomer Hipparchus, and one set of his astronomical observations has been preserved. Although the evidence is not conclusive, it seems that while he was in Athens, he contributed two vital pieces to the mosaic that is Euclid’s *Elements*.

The Eudoxan definition of proportion. The first piece of the *Elements* contributed by Eudoxus was the solution of the problem of incommensurables. This solution is attributed to him on the basis of two facts: (1) Proclus’ comment that Euclid

“arranged many of the theorems of Eudoxus”; (2) an anonymous scholium (commentary) on Euclid’s Book 5, which asserts that the book is the creation “of a certain Eudoxus, [the student] of the teacher Plato” (Allman, 1889, p. 132).

His central observation is a very simple one: Suppose that D and S are respectively the diagonal and side of a square (or pentagon). Even though there are no integers m and n such that $mD = nS$, so that the ratio $D : S$ cannot be defined as $n : m$ for any integers, it remains true that for *every* pair of integers m and n there is a trichotomy: Either $mD < nS$ or $mD = nS$ or $mD > nS$. That fact makes it possible at least to define what is meant by saying that the ratio of D to S is the same for all squares. We simply define the proportion $D_1 : S_1 :: D_2 : S_2$ for two different squares to mean that whatever relation holds between mD_1 and nS_1 for a given pair of integers m and n , that same relation holds between mD_2 and nS_2 . Accordingly, as defined by Euclid at the beginning of Book 5, “A relation that two magnitudes of the same kind have due to their sizes is a *ratio*.” As a definition, this statement is somewhat lacking, but we may paraphrase it as follows: “the relative size of one magnitude in terms of a second magnitude of the same kind is the *ratio* of the first to the second.” We think of size as resulting from measurement and relative size as the result of *dividing* one measurement by another, but Euclid keeps silent on both of these points. Then, “Two magnitudes are said to have a ratio to each other if they are capable of exceeding each other when multiplied.” That is, some multiple of each is larger than the other. Thus, the periphery of a circle and its diameter can have a ratio, but the periphery of a circle and its center cannot. Although the definition of ratio would be hard to use, fortunately there is no need to use it. What is needed is equality of ratios, that is, proportion. That definition follows from the trichotomy just mentioned. Here is the definition given in Book 5 of Euclid, with the material in brackets added from the discussion just given to clarify the meaning:

Magnitudes are said to be in the same ratio, the first to the second [$D_1 : S_1$] and the third to the fourth [$D_2 : S_2$], when, if any equimultiples whatever be taken of the first and third [mD_1 and mD_2] and any equimultiples whatever of the second and fourth [nS_1 and nS_2], the former equimultiples alike exceed, are alike equal to, or are alike less than the latter equimultiples taken in corresponding order [that is, $mD_1 > nS_1$ and $mD_2 > nS_2$, or $mD_1 = nS_1$ and $mD_2 = nS_2$, or $mD_1 < nS_1$ and $mD_2 < nS_2$].

Let us now look again at our conjectured early Pythagorean proof of Euclid’s Proposition 1 of Book 6 of the *Elements*. How much change is required to make this proof rigorous? Very little. Where we have assumed that $3BC = 2CD$, it is only necessary to consider the cases $3BC > 2CD$ and $3BC < 2CD$ and show with the same figure that $3ABC > 2ACD$ and $3ABC < 2ACD$ respectively, and that is done by using the trivial corollary of Proposition 38 of Book 1: *If two triangles have equal altitudes and unequal bases, the one with the larger base is larger*. Eudoxus has not only shown how proportion can be defined so as to apply to incommensurables, he has done so in a way that fits together seamlessly with earlier proofs that apply only in the commensurable case. If only the fixes for bugs in modern computer programs were so simple and effective!

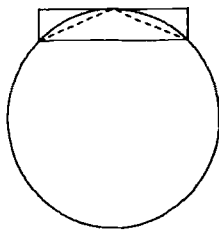


FIGURE 13. The basis of the method of exhaustion.

The method of exhaustion. Eudoxus' second contribution is of equal importance with the first; it is the proof technique known as the *method of exhaustion*. This method is used by both Euclid and Archimedes to establish theorems about areas and solids bounded by curved lines and surfaces. As in the case of the definition of proportion for incommensurable magnitudes, the evidence that Eudoxus deserves the credit for this technique is not conclusive. In his commentary on Aristotle's *Physics*, Simplicius credits the Sophist Antiphon (480–411) with inscribing a polygon in a circle, then repeatedly doubling the number of sides in order to square the circle. However, the perfected method seems to belong to Eudoxus. Archimedes says in the cover letter accompanying his treatise on the sphere and cylinder that it was Eudoxus who proved that a pyramid is one-third of a prism on the same base with the same altitude and that a cone is one-third of the cylinder on the same base with the same altitude. What Archimedes meant by proof we know: He meant proof that meets Euclidean standards, and that can be achieved for the cone only by the method of exhaustion. Like the definition of proportion, the basis of the method of exhaustion is a simple observation: When the number of sides in an inscribed polygon is doubled, the excess of the circle over the polygon is reduced by more than half, as one can easily see from Fig. 13. This observation works together with the theorem that *if two magnitudes have a ratio* and more than half of the larger is removed, then more than half of what remains is removed, and this process continues, then at some point what remains will be less than the smaller of the original two magnitudes (*Elements*, Book 10, Proposition 1). This principle is usually called *Archimedes' principle* because of the frequent use he made of it. The phrase *if two magnitudes have a ratio* is critical, because Euclid's proof of the principle depends on converting the problem to a problem about integers. Since some multiple (n) of the smaller magnitude exceeds the larger, it is only a matter of showing that a finite sequence a_1, a_2, \dots in which each term is less than half of the preceding will eventually reach a point where the ratio $a_k : a_1$ is less than $1/n$.

The definition of ratio and proportion allowed Eudoxus/Euclid to establish all the standard facts about the theory of proportion, including the important fact that similar polygons are proportional to the squares on their sides (*Elements*, Book 6, Propositions 19 and 20). Once that result is achieved, the method of exhaustion makes it possible to establish rigorously what the Pythagoreans had long believed: that similar curvilinear regions are proportional to the squares on similarly situated chords. In particular, it made it possible to prove the fundamental fact that was being used by Hippocrates much earlier: Circles are proportional to the squares on their diameters. This fact is now stated as Proposition 2 of Book 12 of the *Elements*, and the proof given by Euclid is illustrated in Fig. 14.

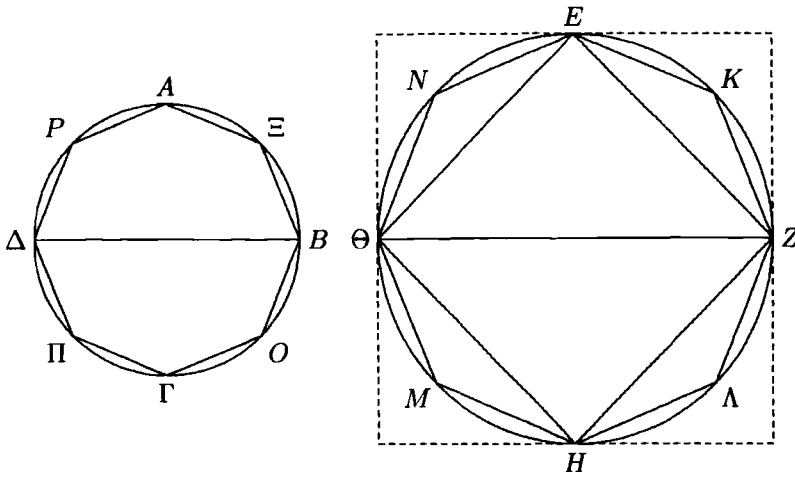


FIGURE 14. Proof that circles are proportional to the squares on their diameters.

Let $AB\Gamma\Delta$ and $EZH\Theta$ be two circles with diameters $B\Delta$ and ΘZ , and suppose that the circles are not proportional to the squares on their diameters. Let the ratio $B\Delta^2 : \Theta Z^2$ be the same as $AB\Gamma\Delta : \Sigma$, where Σ is an area larger or smaller than $EZH\Theta$. Suppose first that Σ is smaller than the circle $EZH\Theta$. Draw the square $EZH\Theta$ inscribed in the circle $EZH\Theta$. Since this square is half of the circumscribed square with sides perpendicular and parallel to the diameter ΘZ , and the circle is smaller than the circumscribed square, the inscribed square is more than half of the circle. Now bisect each of the arcs EZ , ZH , $H\Theta$, and ΘE at points K , Λ , M , and N , and join the polygon $EKZ\Lambda H\Theta NE$. As shown above, doing so produces a larger polygon, and the excess of the circle over this polygon is less than half of its excess over the inscribed square. If this process is continued enough times, the excess of the circle over the polygon will eventually be less than its excess over Σ , and therefore the polygon will be larger than Σ . For definiteness, Euclid assumes that this polygon is the one reached at the first doubling: $EKZ\Lambda H\Theta NE$. In the first circle $AB\Gamma\Delta$, inscribe a polygon $A\Xi B O \Gamma \Pi \Delta P$ similar to $EKZ\Lambda H\Theta NE$. Now the square on $B\Delta$ is to the square on $Z\Theta$ as $A\Xi B O \Gamma \Pi \Delta P$ is to $EKZ\Lambda H\Theta NE$. But also the square on $B\Delta$ is to the square on $Z\Theta$ as the circle $AB\Gamma\Delta$ is to Σ . It follows that $A\Xi B O \Gamma \Pi \Delta P$ is to $EKZ\Lambda H\Theta NE$ as the circle $AB\Gamma\Delta$ is to Σ . Since the circle $AB\Gamma\Delta$ is larger than $A\Xi B O \Gamma \Pi \Delta P$, it follows that Σ must be larger than $EKZ\Lambda H\Theta NE$. But by construction, it is smaller, which is impossible. A similar argument shows that it is impossible for Σ to be larger than $EZH\Theta$.

A look ahead. Ratios as defined by Euclid are always between two magnitudes of the same type. He never considered what we call density, for example, which is the ratio of a mass to a volume. Being always between two magnitudes of the same type, ratios are “dimensionless” in our terms, and could be used as numbers, if only they could be added and multiplied. However, the Greeks obviously did not think of operations on ratios as being the same thing they could do with numbers. In terms of adding, Euclid does say (Book 6, Proposition 24) that if two proportions have the same second and fourth terms, then their first terms and third terms can

be added (first to first and third to third), that is, if $a : b :: c : d$ and $e : b :: f : d$, then $(a+e) : b :: (c+f) : d$. But he did not think of the second and fourth terms in a proportion as denominators or try to get a common denominator. For multiplication of ratios, Euclid gives three separate definitions. In Book 5, Definition 9, he defines the duplicate (which we would call the square) of the ratio $a : b$ to be the ratio $a : c$ if b is the mean proportional between a and c , that is, $a : b :: b : c$. Similarly, when there are four terms in proportion, as in the problem of two mean proportionals, so that $a : b :: b : c :: c : d$, he calls the ratio $a : d$ the triplicate of $a : b$. We would call it the cube of this ratio. Not until Book 6, Definition 5 is there any kind of general definition of the product of two ratios. Even that definition is not in all manuscripts and is believed to be a later interpolation. It goes as follows: *A ratio is said to be the composite of two ratios when the sizes in the two ratios produce something when multiplied by themselves.*¹⁴ This rather vague definition is made worse by the fact that the word for *composed* (*sygkeímena*) is simply a general word for *combined*. It means literally *lying together* and is the same word used when two lines are placed end to end to form a longer line. In that context it corresponds to addition, whereas in the present one it corresponds (but only very loosely) to multiplication. It can be understood only by seeing the way that Euclid operates with it. Given four lines a , b , c , and d , to form the compound ratio $a : b.c : d$, Euclid first takes any two lines m and n such that $a : b :: m : n$. He then finds a line p such that $n : p :: c : d$ and *defines* the compound ratio $a : b.c : d$ to be $m : p$.

There is some arbitrariness in this procedure, since m could be any line. A modern mathematician looking at this proof would note that Euclid could have shortened the labor by taking $m = a$ and $n = b$. The same mathematician would add that Euclid ought to have shown that the final ratio is the same independently of the choice of m , which he did not do. But one must remember that the scholarly community around Euclid was much more intimate than in today's world; he did not have to write a "self-contained" book. In the present instance a glance at Euclid's *Data* shows that he knew what he was doing. The first proposition in that book says that "if two magnitudes A and B are given, then their ratio is given." In modern language, any quantity can be replaced by an equal quantity in a ratio without changing the ratio. The proof is that if $A = \Gamma$ and $B = \Delta$, then $A : \Gamma :: B : \Delta$, and hence by Proposition 16 of Book 5 of the *Elements*, $A : B :: \Gamma : \Delta$. The second proposition of the *Data* draws the corollary that if a given magnitude has a given ratio to a second magnitude, then the second magnitude is also given. That is, if two quantities have the same ratio to a given quantity, then they are equal. From these principles, Euclid could see that the final ratio $m : p$ is what mathematicians now call "well-defined," that is, independent of the initial choice of m .¹⁵ The first use made of this process is in Proposition 23 of Book 6, which asserts that equiangular parallelograms are in the compound ratio of their (corresponding) sides.

With the departure of Eudoxus for Cnidus, we can bring to a close our discussion of Plato's influence on mathematics. If relations between Plato and Eudoxus were less than intimate, as Diogenes Laertius implies, Eudoxus may have drawn off some of Plato's students whose interests were more scientific (in modern terms)

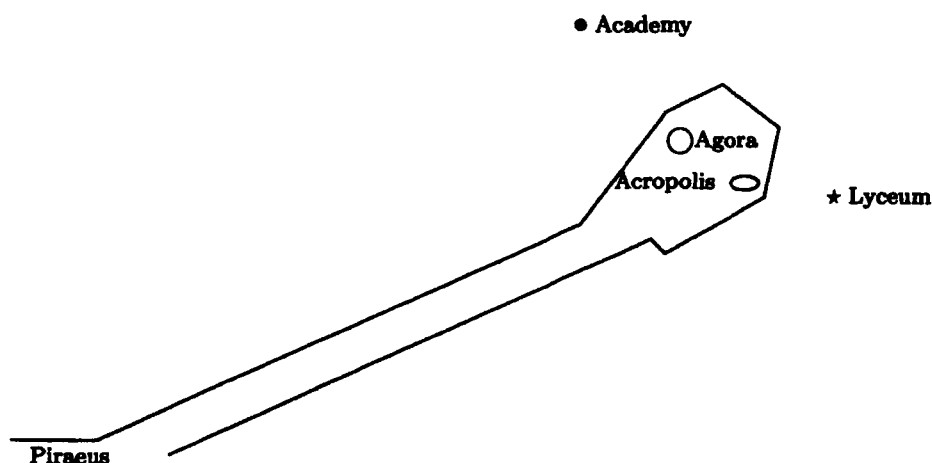
¹⁴ I am aware that the word "in" here is not a literal translation, since the Greek has the genitive case—the sizes *of* the two ratios. But I take *of* here to mean *belonging to*, which is one of the meanings of the genitive case.

¹⁵ A good exposition of the purpose of Euclid's *Data* and its relation to the *Elements* was given by Il'ina (2002), elaborating a thesis of I. G. Bashmakova.

and less philosophical. It is likely that even Plato realized that his attempt to understand the universe through his Forms was not going to work. His late dialogue *Parmenides* gives evidence of a serious rethinking of this doctrine. In any case, it is clear that Plato could not completely dominate the intellectual life of his day.

1.9. Aristotle. Plato died in 347 BCE, and his place as the pre-eminent scholar of Athens was taken a decade after his death by his former pupil Aristotle (384–322 BCE). Aristotle became a student at the Academy at the age of 18 and remained there for 20 years. After the death of Plato he left Athens, traveled, got married, and in 343 became tutor to the future Macedonian King Alexander (the Great), who was 13 years old when Aristotle began to teach him and 16 when he became king on the death of his father. In 335 Aristotle set up his own school, located in the Lyceum, over the hill from the Academy. For the next 12 years he lived and wrote there, producing an enormous volume of speculation on a wide variety of subjects, scientific, literary, and philosophical. In 322 Alexander died, and the Athenians he had conquered turned against his friends. Unlike Socrates, Aristotle felt no obligation to be a martyr to the laws of the polis. He fled to escape the persecution, but died the following year. Aristotle's writing style resembles very much that of a modern scholar, except for the absence of footnotes. Like Plato, in mathematics he seems more like a well-informed generalist than a specialist.

The drive toward the logical organization of science reached its full extent in the treatises of Aristotle. He analyzed reason itself and gave a very thorough and rigorous discussion of formal inference and the validity of various kinds of arguments in his treatise *Prior Analytics*, which was written near the end of his time at the Academy, around 350 BCE. It is easy to picture debates at the Academy, with the mathematicians providing examples of their reasoning, which the logician Aristotle examined and criticized in order to distill his rules for making inferences. In this treatise Aristotle discusses subjects, predicates, and syllogisms connecting the two, occasionally giving a glimpse of some mathematics that may indicate what the mathematicians were doing at the time.



Athens in the fourth century BCE: the waterfront (Piraeus), Academy, and Lyceum.

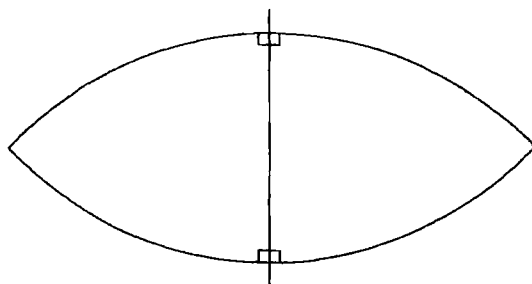


FIGURE 15. How do we exclude the possibility that two lines perpendicular to the same line may intersect each other?

In Book 1 Aristotle describes how to organize the study of a subject, looking for all the attributes and subjects of both of the terms that are to appear in a syllogism. The subject-attribute relation is mirrored in modern thought by the notion of elements belonging to a set. The element is the subject, and the set it belongs to is defined by attributes that can be predicated of all of its elements and no others. Just as sets can be elements of other sets, Aristotle said that the same object can be both a subject and a predicate. He thought, however, that there were some absolute subjects (individual people, for example) that were not predicates of anything and some absolute predicates (what we call abstractions, such as beauty) that were never the subject of any proposition.¹⁶ Aristotle says that the postulates appropriate to each subject must come from experience. If we are thorough enough in stating all the attributes of the fundamental terms in a subject, it will be possible to prove certain things and state clearly what must be assumed.

In Book 2 he discusses ways in which reasoning can go wrong, including the familiar fallacy of “begging the question” by assuming what is to be proved. In this context he offers as an example the people who claim to construct parallel lines. According to him, they are begging the question, starting from premises that cannot be proved without the assumption that parallel lines exist. We may infer that there were around him people who did claim to show how to construct parallel lines, but that he was not convinced. It seems obvious that two lines perpendicular to the same line are parallel, but surely that fact, so obvious to us, would also be obvious to Aristotle. Therefore, he must have looked beyond the obvious and realized that the existence of parallel lines does *not* follow from the immediate properties of lines, circles, and angles. Only when this realization dawns is it possible to see the fallacy in what appears to be common sense. Common sense—that is, human intuition—suggests what can be proved: If two perpendiculars to the same line meet on one side of the line, then they must meet on the other side also, as in Fig. 15. Indeed, Ptolemy did prove this, according to Proclus. But Ptolemy then concluded that two lines perpendicular to the same line cannot meet at all. “But,” Aristotle would have objected, “you have not proved that two lines cannot meet in two different points.” And he would have been right: the assumptions that two lines can meet in only one point and that the two sides of a line are different regions (not connected to each other) are equivalent to assuming that parallel lines exist.

¹⁶ In modern set theory it is necessary to assume that one cannot form an infinite chain of sets a , b , c , . . . such that $b \in a$, $c \in b$, That is, at the bottom of any particular element of a set, there is an “atom” that has no elements.

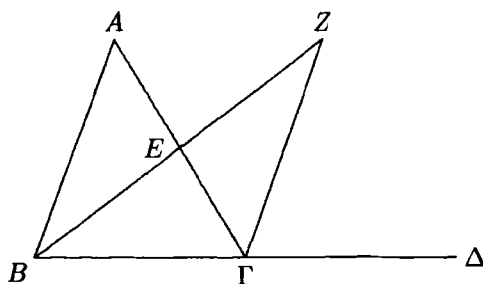


FIGURE 16. The exterior angle theorem.

Euclid deals with this issue in the *Elements* by stating as the last of his assumptions that “two straight lines do not enclose an area.” Oddly, however, he seems unaware of the need for this assumption when proving the main lemma (Book 1, Proposition 16) needed to prove the existence of parallel lines.¹⁷ This proposition asserts that an exterior angle of a triangle is larger than either of the opposite interior angles. Euclid’s proof is based on Fig. 16, in which a triangle $AB\Gamma$ is given with side $B\Gamma$ extended to Δ , forming the exterior angle $A\Gamma\Delta$. He wishes to prove that this angle is larger than the angle at A . To do so, he bisects $A\Gamma$ at E , draws AE , and extends it to Z so that $EZ = AE$. When $Z\Gamma$ is joined, it is seen that the triangles ABE and ΓZE are congruent by the side-angle-side criterion. It follows that the angle at A equals $\angle EZ\Gamma$, which is smaller than $\angle E\Gamma\Delta$, being only a part of it.

In the proof Euclid assumes that the points E and Z are on the same side of line $B\Gamma$. But that is obvious only for triangles small enough to see. It needs to be proved. To be sure, Euclid could have proved it by arguing that if E and Z were on opposite sides of $B\Gamma$, then EZ would have to intersect either $B\Gamma$ or its extension in some point H , and then the line BH passing through Γ and the line BEH would enclose an area. But he did not do that. In fact, the only place where Euclid invokes the assumption that two lines cannot enclose an area is in the proof of the side-angle-side criterion for congruence (Book 1, Proposition 4).¹⁸

Granting that Aristotle was right about this point, we still must wonder why he considered the existence of parallel lines to be in need of proof. Why would he have doubts about something that is so clear on an intuitive level? One possible reason is that parallelism involves the infinite: Parallelism involves the concept that two finite line segments will *never* meet, no matter how far they are extended. If geometry is interpreted physically (say, by regarding a straight line as the path of a light ray), we really have no assurance whatever that parallel lines exist—how could anyone assert with confidence what will happen if two apparently parallel lines are extended to a length of hundreds of light years?

¹⁷ In standard editions of Euclid, there are 14 assumptions, but three of them, concerned with adding equals to equals, doubling equals, and halving equals, are not found in some manuscripts. Gray (1989, p. 46) notes that the fourteenth assumption may be an interpolation by the Muslim mathematician al-Nayrizi, the result of speculation on the foundations of geometry. That would explain its absence from the proof of Proposition 16.

¹⁸ This proof also uses some terms and some hidden assumptions that are visually obvious but which mathematicians nowadays do not allow.

As Aristotle's discussion of begging the question continues, further evidence comes to light that this matter of parallel lines was being debated around 350, and proofs of the existence of parallel lines (Book 1, Proposition 27 of the *Elements*) were being proposed, based on the exterior-angle principle. In pointing out that different false assumptions may lead to the same wrong conclusion, Aristotle notes in particular that the nonexistence of parallel lines would follow if an internal angle of a triangle could be greater than an external angle (not adjacent to it), and also if the angles of a triangle added to more than two right angles.¹⁹ One is almost tempted to say that the mathematicians who analyzed the matter in this way foresaw the non-Euclidean geometry of Riemann, but of course that could not be. Those mathematicians were examining what must be assumed in order to get parallel lines into their geometry. They were not exploring a geometry without parallel lines.

2. Euclid

In retrospect the third century BCE looks like the high-water mark of Greek geometry. Beginning with the *Elements* of Euclid around 300 BCE, this century saw the creation of sublime mathematics in the treatises of Archimedes and Apollonius. It is very tempting to regard Greek geometry as essentially finished after Apollonius, to see everything that came before as leading up to these creations and everything that came after as "polishing up." And indeed, although there were some bright spots afterward and some interesting innovations, none had the scope or the profundity of the work done by these three geometers.

The first of the three major figures from this period is Euclid, who is world famous for his *Elements*, which we have in essence already discussed. This work is so famous, and dominated all teaching in geometry throughout much of the world for so long, that the man and his work have essentially merged. For centuries people said not that they were studying geometry, but that they were studying Euclid. This one work has eclipsed both Euclid's other books and his biography. He did write other books, and two of them—the *Data* and *Optics*—still exist. Others—the *Phænomena*, *Loci*, *Conics*, and *Porisms*—are mentioned by Pappus, who quotes theorems from them.

Euclid is defined for us as the author of the *Elements*. Apart from his writings, we know only that he worked at Alexandria in Egypt just after the death of Alexander the Great. In a possibly spurious passage in Book 7 of his *Synagōgē*, Pappus gives a brief description of Euclid as the most modest of men, a man who was precise but not boastful, like (he implies) Apollonius.

2.1. The *Elements*. As for the *Elements* themselves, the editions that we now have came to us through many hands, and some passages seem to have been added by hands other than Euclid's, especially Theon of Alexandria. We should remember, of course, that Theon was not interested in preserving an ancient literary artifact unchanged; he was trying to produce a good, usable treatise on geometry. Some manuscripts have 15 books, but the last two have since been declared spurious by the experts, so that the currently standard edition has 13 books, the last of which looks suspiciously less formal than the first 12, leading some to doubt that Euclid wrote it. Leaving aside the thorny question of which parts were actually written

¹⁹ Field and Gray (1987, p. 64) note that this point has been made by many authors since Aristotle.

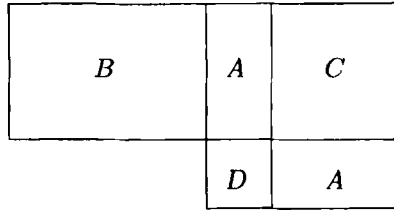


FIGURE 17. Expression of a rectangle as the difference of two squares.

by Euclid, we give just a summary description of the contents, since we have seen them coming together in the work of the Pythagoreans, Plato, and Aristotle.

The contents of the first book of the *Elements* are covered in the standard geometry courses given in high schools. This material involves the elementary geometric constructions of copying angles and line segments, drawing squares, and the like and the basic properties of parallelograms, culminating in the Pythagorean theorem (Proposition 47). In addition, these properties are applied to the problem of transformation of areas, leading to the construction of a parallelogram with a given base angle, and equal in area to any given polygon (Proposition 45). There the matter rests until the end of Book 2, where it is shown (Proposition 14) how to construct a square equal to any given polygon.

Book 2 contains geometric constructions needed to solve problems that may involve quadratic incommensurables without resorting to the Eudoxan theory of proportion. For example, a fundamental result is Proposition 5: *If a straight line is cut into equal and unequal segments, the rectangle contained by the unequal segments of the whole together with the square on the straight line between the points of the section is equal to the square on the half.* This proposition is easily seen using Fig. 17, in terms of which it asserts that $(A+B)+D=2A+C+D$; that is, $B=A+C$.

This proposition, in arithmetic form, appeared as a fundamental tool in the cuneiform tablets. For if the unequal segments of the line are regarded as two unknown quantities, then half of the segment is precisely their average, and the straight line between the points (that is, the segment between the midpoint of the whole segment and the point dividing the whole segment into unequal parts) is precisely what we called earlier the semidifference. Thus, this proposition says that the square of the average equals the product plus the square of the semidifference; and that result was fundamental for solving the important problems of finding two numbers given their sum and product or their difference and product. However, those geometric constructions do not appear until Book 6. These application problems could have been solved in Book 2 in the case when the excess or defect is a square. Instead, these special cases were passed over and the more general results, which depend on the theory of proportion, were included in Book 6.

Book 2 also contains the construction of what came to be known as the *Section*, that is, the division of a line in mean and extreme ratio so that the whole is to one part as that part is to the other. But Euclid is not ready to prove that version yet, since he doesn't have the theory of proportion. Instead, he gives what must have been the original form of this proposition (Proposition 11): *to cut a line so that the rectangle on the whole and one of the parts equals the square on the other*

part. After it is established that four lines are proportional when the rectangle on the means equals the rectangle on the extremes (Proposition 16, Book 6), it becomes possible to convert this construction into the construction of the Section (Proposition 30, Book 6).

Books 3 and 4 take up topics familiar from high-school geometry: circles, tangents and secants, and inscribed and circumscribed polygons. In particular, Book 4 shows how to inscribe a regular pentagon in a circle (Proposition 11) and how to circumscribe a regular pentagon about a circle (Proposition 12), then reverses the figures and shows how to get the circles given the pentagon (Propositions 13 and 14). After the easy construction of a regular hexagon (Proposition 15), Euclid finishes off Book 4 with the construction of a regular pentakaidecagon (15-sided polygon, Proposition 16).

Book 5 contains the Eudoxan theory of geometric proportion, in particular the construction of the mean proportional between two lines (Proposition 13). In Book 6 this theory is applied to solve the problems of application with defect and excess. A special case of the latter, in which it is required to construct a rectangle on a given line having area equal to the square on the line and with a square excess is the very famous Section (Proposition 30). Euclid phrases the problem as follows: *to divide a line into mean and extreme ratio*. This means to find a point on the line so that the whole line is to one part as that part is to the second part. The Pythagorean theorem is then generalized to cover not merely the squares on the sides of a right triangle, but any similar polygons on those sides (Proposition 31). The book finishes with the well-known statement that central and inscribed angles in a circle are proportional to the arcs they subtend.

Books 7-9 were discussed in Chapter 7. They are devoted to Pythagorean number theory. Here, since irrationals cannot occur, the notion of proportion is redefined to eliminate the need for the Eudoxan technique.

Book 10 occupies fully one-fourth of the entire length of the *Elements*. For its sheer bulk, one would be inclined to consider it the most important of all the 13 books, yet its 115 propositions are among the least studied of all, principally because of their technical nature. The irrationals constructed in this book by taking square roots are needed in the theory developed in Book 13 for inscribing regular solids in a sphere (that is, finding the lengths of their sides knowing the radius of the sphere). The book begins with the operating principle of the method of exhaustion, also known as the principle of Archimedes. The way to demonstrate incommensurability through the Euclidean algorithm then follows as Proposition 2: *If, when the smaller of two given quantities is continually subtracted from the larger, that which is left never divides evenly the one before it, the quantities are incommensurable*. We used this method of showing that the side and diagonal of a regular pentagon are incommensurable in Chapter 8.

Book 11 contains the basic parts of the solid geometry of planes, parallelepipeds, and pyramids. The theory of proportion for these solid figures is developed in Book 12, where one finds neatly tucked away the theorem that circles are proportional to the squares on their diameters (Proposition 2), which we quoted above.

Book 12 continues the development of solid geometry by establishing the usual proportions and volume relations for solid figures; for example, a triangular prism can be divided by planes into three pyramids, all having the same volume (Proposition 7), a cone has one-third the volume of a cylinder on the same base, similar

cones and cylinders are proportional to the cubes of their linear dimensions, ending with the proof that spheres are proportional to the cubes on their diameters (Proposition 18). As we noted above, Archimedes (or someone who edited his works) credited these theorems to Eudoxus.

Book 13, the last book of the *Elements*, is devoted to the construction of the regular solids and the relation between their dimensions and the dimensions of the sphere in which they are inscribed. The last proposition (Proposition 18) sets out the sides of these regular solids and their ratios to one another. An informal discussion following this proposition concludes that there can be only five regular solids.

2.2. The *Data*. Euclid's *Elements* assume a certain familiarity with the principles of geometric reasoning, principles that are explained in more detail in the *Data*. The Greek name of this work (*Dedómena*) means [Things That Are] *Given*, just as *Data* does in Latin. The propositions in this book can be interpreted in various ways. Some can be looked at as *uniqueness* theorems. For example (Proposition 53), if the shapes—that is, the angles and ratios of the sides—are given for two polygons, and the ratio of the areas of the polygons is given, then the ratio of any side of one to any side of the other is given. Here, being given means being uniquely determined. Uniqueness is needed in proofs and constructions so that one can be sure that the result will be the same no matter what choices are made. It is an issue that arises frequently in modern mathematics, where operations on sets are defined by choosing representatives of the sets; when that is done, it is necessary to verify that the operation is *well defined*, that is, independent of the choice made. In geometry we frequently say, “Let ABC be a triangle having the given properties *and having such-and-such a property*,” such as being located in a particular position. In such cases we need to be sure that the additional condition does not restrict the generality of the argument. In another sense, this same proposition reassures the reader that an explicit construction is *possible*, and removes the necessity of including it in the exposition of a theorem.

Other propositions assert that certain properties are *invariant*. For example (Proposition 81), when four lines A , B , Γ , and Δ are given, and the line H is such that $\Delta : E = A : H$, where E is the fourth proportional to A , B , and Γ , then $\Delta : \Gamma = B : H$. This last proposition is a lemma that can be useful in working out locus problems, which require finding a curve on which a point must lie if it satisfies certain prescribed conditions. Finally, a modern mathematician might interpret the assertion that an object is “given” as saying that the object “exists” and can be meaningfully talked about. To Euclid, that existence would mean that the object was explicitly constructible.

3. Archimedes

Archimedes is one of a small number of mathematicians of antiquity of whose works we know more than a few fragments and of whose life we know more than the approximate time and place. The man indirectly responsible for his death, the Roman general Marcellus, is also indirectly responsible for the preservation of some of what we know about him. Archimedes lived in the Greek city of Syracuse on the island of Sicily during the third century BCE and is said by Plutarch to have been “a relative and a friend” of King Hieron II. Since Sicily lies nearly on a direct line between Carthage and Rome, it became embroiled in the Second Punic War.

Marcellus took the city of Syracuse after a long siege, and Archimedes was killed by a Roman soldier in the chaos of the final fall of the city. In the course of writing a biography of Marcellus, the polymath Plutarch included some information on mathematics and philosophy in general.

According to Plutarch's biography of Marcellus, the general was very upset that Archimedes had been killed and had his body buried in a suitably imposing tomb. According to Eutocius, a biography of Archimedes was written by a certain Heracleides, who is mentioned in some of Archimedes' letters. However, no copy of this biography is known to exist today.

There are many legends connected with Archimedes, scattered among the various sources. Plutarch, for instance, says that Archimedes made many mechanical contrivances but generally despised such work in comparison with pure mathematical thought. Plutarch also reports three different stories of the death of Archimedes and tells us that Archimedes wished to have a sphere inscribed in a cylinder carved on his tombstone. The famous story that Archimedes ran naked through the streets shouting "Eureka!" ("I've got it!") when he discovered the principle of specific gravity in the baths is reported by the Roman architect Vitruvius. Proclus gives another well-known anecdote: that Archimedes built a system of pulleys that enabled him (or King Hieron) single-handedly to pull a ship through the water. Finally, Plutarch and Pappus both quote Archimedes as saying in connection with his discovery of the principle of the lever that if there were another Earth, he could move this one by standing on it.

With Archimedes we encounter the first author of a considerable body of original mathematical research that has been preserved to the present day. He was one of the most versatile, profound, creative, imaginative, rigorous, and influential mathematicians who ever lived. Ten of Archimedes' treatises have come down to the present, along with a *Book of Lemmas* that seems to be Archimedean. Some of these works are prefaced by a "cover letter" intended to explain their contents to the person to whom Archimedes sent them. These correspondents of Archimedes were: Gelon, son of Hieron II and one of the kings of Syracuse during Archimedes' life; Dositheus, a student of Archimedes' student and close friend Conon; and Eratosthenes, an astronomer who worked in Alexandria. Like the manuscripts of Euclid, all of the Archimedean manuscripts date from the ninth century or later. These manuscripts have been translated into English and published by various authors. A complete set of Medieval manuscripts of Archimedes' work has been published by Marshall Clagett in the University of Wisconsin series on Medieval Science.

The 10 treatises referred to above are the following.

1. *On the Equilibrium of Planes*, Part I
2. *Quadrature of the Parabola*
3. *On the Equilibrium of Planes*, Part II
4. *On the Sphere and the Cylinder*, Parts I and II
5. *On Spirals*
6. *On Conoids and Spheroids*
7. *On Floating Bodies*
8. *Measurement of a Circle*
9. *The Sand-reckoner*
10. *The Method*

References by Archimedes himself and other mathematicians tell of the existence of other works by Archimedes of which no manuscripts are now known to exist. These include works on the theory of balances and levers, optics, the regular polyhedra, the calendar, and the construction of mechanical representations of the motion of heavenly bodies. In 1998 a palimpsest of Archimedes' work was sold at auction for \$2 million (see Plate 6).

From this list we can see the versatility of Archimedes. His treatises on the equilibrium of planes and floating bodies contain principles that are now fundamental in mechanics and hydrostatics. The works on the quadrature of the parabola, conoids and spheroids, the measurement of the circle, and the sphere and cylinder extend the theory of proportion, area, and volume found in Euclid for polyhedra and polygons to the more complicated figures bounded by curved lines and surfaces. The work on spirals introduces a new class of curves, and develops the theory of length, area, and proportion for them.

Since we do not have space to discuss all of Archimedes' geometry, we shall confine our discussion to what may be his greatest achievement: finding the surface area of a sphere. In addition, because of its impact on the issues involving proof that we have been discussing, we shall discuss his *Method*.

3.1. The area of a sphere. Archimedes' two works on the sphere and cylinder were sent to Dositheus. In the letter accompanying the first of these he gives some of the history of the problem. Archimedes considered his results on the sphere to be rigorously established, but he did have one regret:

It would have been beneficial to publish these results when Conon was alive, for he is the one we regard as most capable of understanding and rendering a proper judgment on them. But, as we think it well to communicate them to the initiates of mathematics, we send them to you, having rewritten the proofs, which those versed in the sciences may scrutinize.

The fact that a pyramid is one-third of a prism on the same base and altitude is Proposition 7 of Book 12 of Euclid's *Elements*. Thus Archimedes could say confidently that this theorem was well established. Archimedes sought the surface area of a sphere by finding the lateral surface area of a frustum of a cone and the lateral area of a right cylinder. In our terms the area of a frustum of a cone with upper radius r , lower radius R , and side of slant height h is $\pi h(R+r)$. Archimedes phrased this fact by saying that the area is that of a circle whose radius is the mean proportional between the slant height and the sum of the two radii; that is, the radius is $\sqrt{h(R+r)}$. Likewise, our formula for the lateral surface area of a cylinder of radius r and height h is $2\pi rh$. Archimedes said it was the area of a circle whose radius is the mean proportional between the diameter and height of the cylinder.

These results can be applied to the figures generated by revolving a circle about a diameter with certain chords drawn. Archimedes showed (Proposition 22) that

$$(BB' + CC' + \cdots + KK' + LM) : AM = A'B : BA$$

in Fig. 18.

This result is easily derived by connecting B' to C , C' to K , and K' to L and considering the ratios of the legs of the resulting similar triangles. These ratios can be added. All that then remains is to cross-multiply this proportion and use the expressions already derived for the area of a frustum of a cone. One finds easily

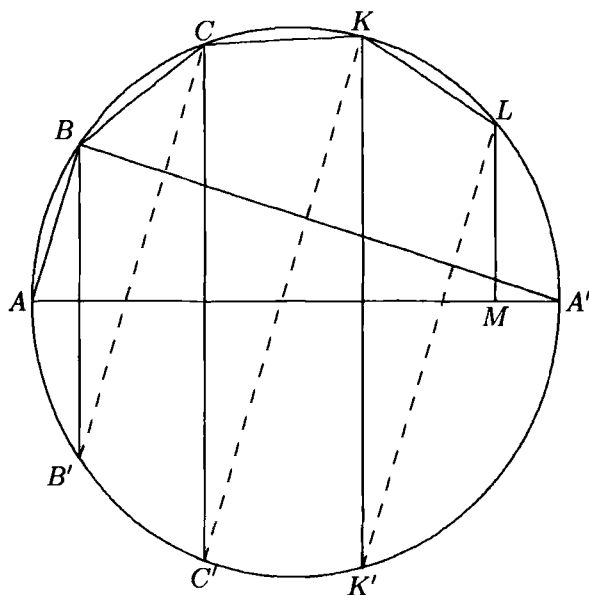


FIGURE 18. Finding the surface area of a sphere.

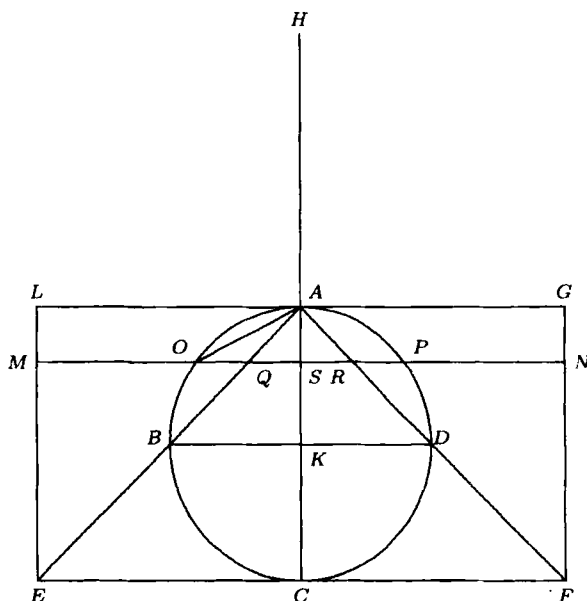
that the area of the surface obtained by revolving the broken line $ABCKL$ about the axis AA' is $\pi AM \cdot A'B$. The method of exhaustion then shows that the product $AM \cdot A'B$ can be made arbitrarily close to the square of AA' ; it therefore gives the following result (Proposition 33): *The surface of any sphere is equal to four times the greatest circle in it.*

By the same method, using the inscribed right circular cone with the equatorial circle of the sphere as a base, Archimedes shows that the volume of the sphere is four times the volume of this cone. He then obtains the relations between the areas and volumes of the sphere and circumscribed closed cylinder. He finishes this first treatise with results on the area and volume of a segment of a sphere, that is, the portion of a sphere cut off by a plane. This argument is the only ancient proof of the area and volume of a sphere that meets Euclidean standards of rigor.

Three remarks should be made on this proof. First, in view of the failure of efforts to square the circle, it seems that the later Greek mathematicians had two standard areas, the circle and the square. Archimedes expressed the area of a sphere in terms of the area of a circle. Second, as we have seen, the volume of a sphere was found in China several centuries after Archimedes' time, but the justification for it involved intuitive principles such as Cavalieri's principle that do not meet Euclidean standards. Third, Archimedes did not *discover* this theorem by Euclidean methods. He told how he came to discover it in his *Method*.

3.2. The Method. Early in the twentieth century the historian of mathematics J. L. Heiberg, reading in a bibliographical journal of 1899 the account of the discovery of a tenth-century manuscript with mathematical content, deduced from a few quotations that the manuscript was a copy²⁰ of a work of Archimedes. In

²⁰ A copy, not Archimedes' own words, since it was written in the Attic dialect, while Archimedes wrote in Doric. It is interesting that in the statement of his first theorem Archimedes refers to a "section of a right-angled cone $AB\Gamma$," and then immediately in the proof says, "since $AB\Gamma$ is a



a cone, the rectangle $LGFE$ generates a cylinder, and each horizontal line such as MN generates a disk. The point A is the midpoint of CH . Archimedes shows that the area of the disk generated by revolving QR plus the area of the disk generated by revolving OP has the same ratio to the area of the disk generated by revolving MN that AS has to AH . It follows from his work on the equilibrium of planes that if the first two of these disks are hung at H , they will balance the third disk about A as a fulcrum. Archimedes concluded that the sphere and cone together placed with their centers of gravity at H would balance (about the point A) the cylinder, whose center of gravity is at K .

Therefore,

$$HA : AK = (\text{cylinder}) : (\text{sphere} + \text{cone}).$$

But $HA = 2AK$. Therefore, the cylinder equals twice the sum of the sphere and the cone AEF . And since it is known that the cylinder is three times the cone AEF , it follows that the cone AEF is twice the sphere. But since $EF = 2BD$, cone AEF is eight times cone ABD , and the sphere is four times the cone ABD .

From this fact Archimedes easily deduces the famous result allegedly depicted on his tombstone: *The cylinder circumscribed about a sphere equals the volume of the sphere plus the volume of a right circular cone inscribed in the cylinder.*

Having concluded the demonstration, Archimedes reveals that this method enabled him to discover the area of a sphere. He writes

For I realized that just as every circle equals a triangle having as its base the circumference of the circle and altitude equal to the [distance] from the center to the circle [that is, the radius], in the same way every sphere is equal to a cone having as its base the surface [area] of the sphere and altitude equal to the [distance] from the center to the sphere.

The *Method* gives an inside view of the route by which Archimedes discovered his results. The method of exhaustion is convincing as a method of proving a theorem, but useless as a way of discovering it. The *Method* shows us Archimedes' route to discovery.

4. Apollonius

From what we have already seen of Greek geometry we can understand how the study of the conic sections came to seem important. From commentators like Pappus we know of treatises on the subject by Aristaeus, a contemporary of Euclid who is said to have written a book on *Solid Loci*, and by Euclid himself. We have also just seen that Archimedes devoted a great deal of attention to the conic sections. The only treatise on the subject that has survived, however, is that of Apollonius, and even for this work, unfortunately, no faithful translation into English exists. The version most accessible is that of Heath, who says in his preface that writing his translation involved "the substitution of a new and uniform notation, the condensation of some propositions, the combination of two or more into one, some slight re-arrangements of order for the purpose of bringing together kindred propositions in cases where their separation was rather a matter of accident than indicative of design, and so on." He might also have mentioned that he supplemented Apollonius' purely synthetic methods with analytic arguments, based on the algebraic notation we are familiar with. All this labor has no doubt made Apollonius more

readable. On the other hand, Apollonius' work is no longer current research, and from the historian's point of view this kind of tinkering with the text only makes it harder to place the work in proper perspective.

In contrast to his older contemporary Archimedes, Apollonius remains a rather obscure figure. His dates can be determined from the commentary written on the *Conics* by Eutocius. Eutocius says that Apollonius lived in the time of the king Ptolemy Euergetes and defends him against a charge by Archimedes' biographer Heracleides that Apollonius plagiarized results of Archimedes. Eutocius' information places Apollonius reliably in the second half of the third century BCE, perhaps a generation or so younger than Archimedes.

Pappus says that as a young man Apollonius studied at Alexandria, where he made the acquaintance of a certain Eudemus. It is probably this Eudemus to whom Apollonius addresses himself in the preface to Book 1 of his treatise. From Apollonius' own words we know that he had been in Alexandria and in Perga, which had a library that rivaled the one in Alexandria. Eutocius reports an earlier writer, Geminus by name, as saying that Apollonius was called "the great geometer" by his contemporaries. He was highly esteemed as a mathematician by later mathematicians, as the quotations from his works by Ptolemy and Pappus attest. In Book 12 of the *Almagest*, Ptolemy attributes to Apollonius a geometric construction for locating the point at which a planet begins to undergo retrograde motion. From these later mathematicians we know the names of several works by Apollonius and have some idea of their contents. However, only two of his works survive to this day, and for them we are indebted to the Islamic mathematicians who continued to work on the problems that Apollonius considered important. Our present knowledge of Apollonius' *Cutting Off of a Ratio*, which contains geometric problems solvable by the methods of application with defect and excess, is based on an Arabic manuscript, no Greek manuscripts having survived. Of the eight books of Apollonius' *Conics*, only seven have survived in Arabic and only four in Greek.

4.1. History of the *Conics*. The evolution of the *Conics* was reported by Pappus five centuries after they were written in Book 7 of his *Collection*.

By filling out Euclid's four books on the conics and adding four others Apollonius produced eight books on the conics. Aristaeus. . . and all those before Apollonius, called the three conic curves sections of acute-angled, right-angled, and obtuse-angled cones. But since all three curves can be produced by cutting any of these three cones, as Apollonius seems to have objected, [noting] that some others before him had discovered that what was called a section of an acute-angled cone could also be [a section of] a right- or obtuse-angled cone. . . changing the nomenclature, he named the so-called acute section an ellipse, the right section a parabola, and the obtuse section a hyperbola.

As already mentioned, the first four books of Apollonius' *Conics* survived in Greek, and seven of the eight books have survived in Arabic; the astronomer Edmund Halley (1656–1743) published a Latin edition of them in 1710.

4.2. Contents of the *Conics*. In a preface addressed to the aforementioned Eudemus, Apollonius lists the important results of his work: the description of the sections, the properties of the figures relating to their diameters, axes, and

asymptotes, things necessary for analyzing problems to see what data permit a solution, and the three- and four-line locus. He continues:

The third book contains many remarkable theorems of use for the construction of solid loci and for distinguishing when problems have a solution, of which the greatest part and the most beautiful are new. And when we had grasped these, we knew that the three-line and four-line locus had not been constructed by Euclid, but only a chance part of it and that not very happily. For it was not possible for this construction to be completed without the additional things found by us.

We have space to discuss only the definition and construction of the conic sections and the four-line locus problem, which Apollonius mentions in the passage just quoted.

4.3. Apollonius' definition of the conic sections. The earlier use of conic sections had been restricted to cutting cones with a plane perpendicular to a generator. As we saw in our earlier discussion, this kind of section is easy to analyze and convenient in the applications for which it was intended. In fact, only one kind of hyperbola, the rectangular, is needed for duplicating the cube and trisecting the angle. The properties of a general section of a general cone were not discussed. Also, it was considered a demerit that the properties of these planar curves had to be derived from three-dimensional figures. Apollonius set out to remove these gaps in the theory.

First it was necessary to define a cone as the figure generated by moving a line around a circle while one of its points, called the *apex* and lying outside the plane of the circle, remains fixed. Next, it was necessary to classify all the sections of a cone that happen to be circles. Obviously, those sections include all sections by planes parallel to the plane of the generating circle (Book 1, Proposition 4). Surprisingly, there is another class of sections that are also circles, called *subcontrary* sections. Once the circles are excluded, the remaining sections must be parabolas, hyperbolas, and ellipses. We have space only to consider Apollonius' construction of the ellipse. His construction of the other conics is very similar. Consider the planar section of a cone in Fig. 20, which cuts all the generators of the cone on the same side of its apex. This condition is equivalent to saying that the cutting intersects both sides of the axial triangle. Apollonius proved that there is a certain line, which he called *the [up]right side*, now known by its Latin name *latus rectum*, such that the square on the ordinate from any point of the section to its axis equals the rectangle applied to the portion of the axis cut off by this ordinate (the abscissa) and whose defect on the axis is similar to the rectangle formed by the axis and the *latus rectum*. He gave a rule, too complicated to go into here, for constructing the *latus rectum*. This line characterized the shape of the curve. Because of its connection with the problem of application with defect, he called the resulting conic section an *ellipse*. Similar connections with the problems of application and application with excess respectively arise in Apollonius' construction of the parabola and hyperbola. These connections motivated the names he gave to these curves.

In Fig. 20 the *latus rectum* is the line EH , and the locus condition is that the square on LM equal the rectangle on EO and EM ; that is, $LM^2 = EO \cdot EM$.

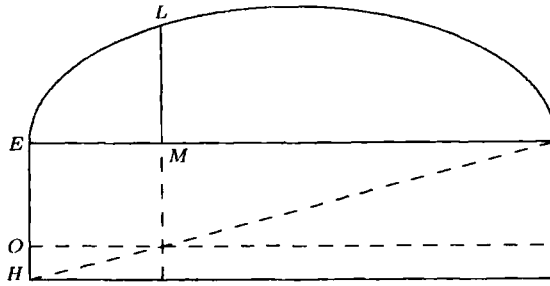


FIGURE 20. Apollonius' construction of the ellipse.

In one sense, this locus definition for an ellipse is not far removed from what we now think of as the equation of the ellipse, but that small gap was unbridgeable in Apollonius' time. If we write $LM = y$ and $EM = x$ in Fig. 20 (so that we are essentially taking rectangular coordinates with origin at E), we see that Apollonius is claiming that $y^2 = x \cdot EO$. Now, however, $EO = EH - OH$, and EH is constant, while OH is directly proportional to EM , that is, to x . Specifically, the ratio of OH to EM is the same as the ratio of EH to the axis. Thus, if we write $OH = kx$ —the one, crucial step that Apollonius could not take, since he did not have the concept of a dimensionless constant of proportionality—and denote the latus rectum EH by C , we find that Apollonius' locus condition can be stated as the equation $y^2 = Cx - kx^2$. By completing the square on x , transposing terms, and dividing by the constant term, we can bring this equation into what we now call the standard form for an ellipse with center at $(a, 0)$:

$$(2) \quad \frac{(x - a)^2}{a^2} + \frac{y^2}{b^2} = 1,$$

where $a = C/(2k)$ and $b = C\sqrt{k}$. In these terms the latus rectum C is $2b^2/a$. Apollonius, however, did *not* have the concept of an equation nor the symbolic algebraic notation we now use, and if he did have, he would still have needed the letter k used above as a constant of proportionality. These “missing” pieces gave his work on conics a ponderous character with which most mathematicians today have little patience.

Apollonius' constructions of the parabola and hyperbola also depend on the latus rectum. He was the first to take account of the fact that a plane that produces a hyperbola must cut both nappes of the cone. He regarded the two branches as two hyperbolas, referring to them as “opposites” and reserving the term *hyperbola* for either branch. For the hyperbola Apollonius proves the existence of *asymptotes*, that is, a pair of lines through the center that never meet the hyperbola but such that any line through the center passing into the region containing the hyperbola does meet the hyperbola. The word *asymptote* means literally *not falling together*, that is, not intersecting.

Books 1 and 2 of the *Conics* are occupied with finding the proportions among line segments cut off by chords and tangents on conic sections, the analogs of results on circles in Books 3 and 4 of Euclid. These constructions involve finding the tangents to the curves satisfying various supplementary conditions such as being parallel to a given line.

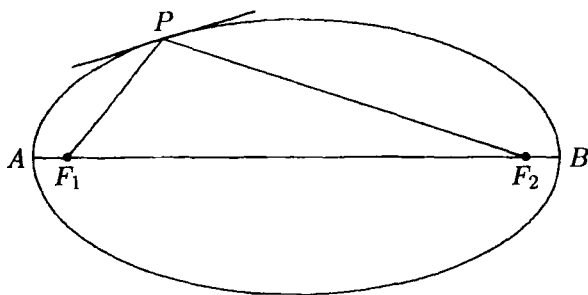


FIGURE 21. Focal properties of an ellipse.

4.4. Foci and the three- and four-line locus. We are nowadays accustomed to constructing the conic sections using the focus-directrix property, so that it comes as a surprise that the original expert on the subject does not seem to recognize the importance of the foci. He never mentions the focus of a parabola, and for the ellipse and hyperbola he refers to these points only as “the points arising out of the application.” The “application” he has in mind is explained in Book 3. Propositions 48 and 52 read as follows:

(Proposition 48) *If in an ellipse a rectangle equal to the fourth part of the figure is applied from both sides to the major axis and deficient by a square figure, and from the points resulting from the application straight lines are drawn to the ellipse, the lines will make equal angles with the tangent at that point.*

(Proposition 52) *If in an ellipse a rectangle equal to the fourth part of the figure is applied from both sides to the major axis and deficient by a square figure, and from the points resulting from the application straight lines are drawn to the ellipse, the two lines will be equal to the axis.*

The “figure” referred to is the rectangle whose sides are the major axis of the ellipse and the latus rectum. In Fig. 21 the points F_1 and F_2 must be chosen on the major axis AB so that $AF_1 \cdot F_1B$ and $AF_2 \cdot BF_2$ both equal one-fourth of the area of the rectangle formed by the axis AB and the latus rectum. Proposition 48 expresses the focal property of these two points: Any ray of light emanating from one will be reflected to the other. Proposition 52 is the *string property* that characterizes the ellipse as the locus of points such that the sum of the distances to the foci is constant. These are just two of the theorems Apollonius called “strange and beautiful.” Apollonius makes little use of these properties, however, and does not discuss the use of the string property to draw an ellipse.

A very influential part of the *Conics* consists of Propositions 54–56 of Book 3, which contain the theorems that Apollonius claimed (in his cover letter) would provide a solution to the three- and four-line locus problems. Both in their own time and because of their subsequent influence, the three- and four-line locus problems have been of great importance for the development of mathematics. These propositions involve the proportions in pieces of chords inscribed in a conic section. Three propositions are needed because the hyperbola requires two separate statements to cover the cases when the points involved lie on different branches of the hyperbola.

Proposition 54 asserts that given a chord $A\Gamma$ such that the tangents at the endpoints meet at Δ , and the line from Δ to the midpoint E of the chord meets the conic at B , any point Θ on the conic has the following property (Fig. 22). *The*

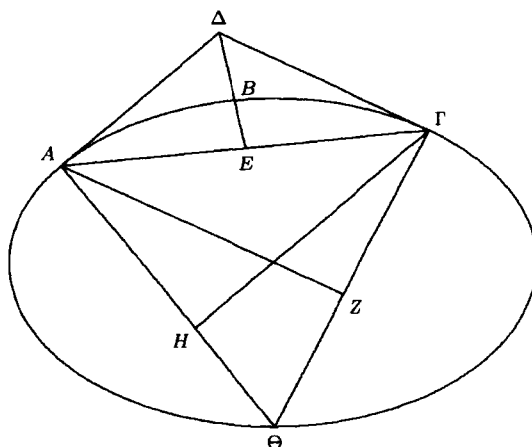


FIGURE 22. The basis for solving the four-line locus problem.

lines from Θ to A and Γ meet the lines through Γ and A respectively, each parallel to the tangent through the other endpoint, in points H and Z respectively such that the following proportion holds: The ratio of the rectangle on AZ and ΓH to the square on $A\Gamma$ is the composite of the ratio of the square on EB to the square on $B\Delta$ and the ratio of the rectangle on $A\Delta$ and $\Delta\Gamma$ to the rectangle on AE and $E\Gamma$. As we would write this relation,

$$AZ \cdot \Gamma H : A\Gamma^2 :: EB^2 : B\Delta^2 \cdot A\Delta \cdot \Delta\Gamma : AE \cdot E\Gamma.$$

It is noteworthy that this theorem involves an expression that seldom occurs in Greek mathematics: a composition of ratios involving squares. If we thought of it in our terms, we would say that Apollonius was working in four dimensions. But he wasn't. The composition of two ratios is possible only when both ratios are between quantities of the same type, in this case areas. The proof given by Apollonius has little in common with the way we would proceed nowadays, by multiplying and dividing. Apollonius had to find an area C such that $AZ \cdot \Gamma H : C :: EB^2 : B\Delta^2$ and $C : A\Gamma^2 :: A\Delta \cdot \Delta\Gamma : ZE \cdot E\Gamma$. He could then "cancel" C in accordance with the definition of compound ratio.

It is not at all obvious how this proposition makes it possible to solve the four-line locus, and Apollonius does not fill in the details. We shall not attempt to do so, either. To avoid excessive complexity, we merely state the four-line locus problem and illustrate it. The data for the problem are four lines, which for definiteness we suppose to intersect two at a time, and four given angles, one corresponding to each line. The problem requires the locus of points P such that if lines are drawn from P to the four lines, each making the corresponding angle with the given line (for simplicity all shown as right angles in Fig. 23), the rectangle on two of the lines will have a constant ratio to the rectangle on the other two. The solution is in general a pair of conics.

The origin of this kind of problem may lie in the problem of two mean proportionals, which was solved by drawing fixed reference lines and finding the loci of points satisfying two conditions resembling this. The square on the line drawn perpendicular to one reference line equals the rectangle on a fixed line and the line

drawn to the other reference line. The commentary on this problem by Pappus, who mentioned that Apollonius had left a great deal unfinished in this area, inspired Fermat and Descartes to take up the implied challenge and solve the problem completely. Descartes offered his success in solving the locus problem to any number of lines as proof of the value of his geometric methods.

Questions and problems

10.1. Show how it would be possible to compute the distance from the center of a square pyramid to the tip of its shadow without entering the pyramid, after first driving a stake into the ground at the point where the shadow tip was located at the moment when vertical poles cast shadows equal to their length.

10.2. Describe a mechanical device to draw the quadratrix of Hippias. You need a smaller circle of radius $2/\pi$ times the radius that is rotating, so that you can use it to wind up a string attached to the moving line; or conversely, you need the rotating radius to be $\pi/2$ times the radius of the circle pulling the line. How could you get such a pair of circles?

10.3. Prove that the problem of constructing a rectangle of prescribed area on part of a given base a in such a way that the defect is a square is precisely the problem of finding two numbers given their sum and product (the two numbers are the lengths of the sides of the rectangle). Similarly, prove that the problem of application with square excess is precisely the problem of finding two numbers (lengths) given their difference and product.

10.4. Show that the problem of application with square excess has a solution for any given area and any given base. What restrictions are needed on the area and base in order for the problem of application with square defect to have a solution?

10.5. Use an argument similar to the argument in Chapter 8 showing that the side and diagonal of a pentagon are incommensurable to show that the side and diagonal of a square are incommensurable. That is, show that the Euclidean algorithm, when applied to the diagonal and side of a square, requires only two steps to produce the side and diagonal of a smaller square, and hence can never produce an equal pair. To do so, refer to Fig. 24.

In this figure $AB = BC$, angle ABC is a right angle, AD is the bisector of angle CAB , and DE is drawn perpendicular to AC . Prove that $BD = DE$, $DE = EC$, and $AB = AE$. Then show that the Euclidean algorithm starting with the pair (AC, AB) leads first to the pair $(AB, EC) = (BC, BD)$, and then to the pair $(CD, BD) = (CD, DE)$, and these last two are the diagonal and side of a square.

10.6. It was stated above that Thales might have used the Pythagorean theorem in order to calculate the distance from the center of the Great Pyramid to the tip of its shadow. How could this distance be computed without the Pythagorean theorem?

10.7. State the paradoxes of Zeno in your own words and tell how you would have advised the Pythagoreans to modify their system in order to avoid these paradoxes.

10.8. Do we share any of the Pythagorean mysticism about geometric shapes that Proclus mentioned? Think of the way in which we refer to an honorable person as *upright*, or speak of getting a *square deal*, while a person who cheats is said to be *crooked*. Are there other geometric images in our speech that have ethical connotations?

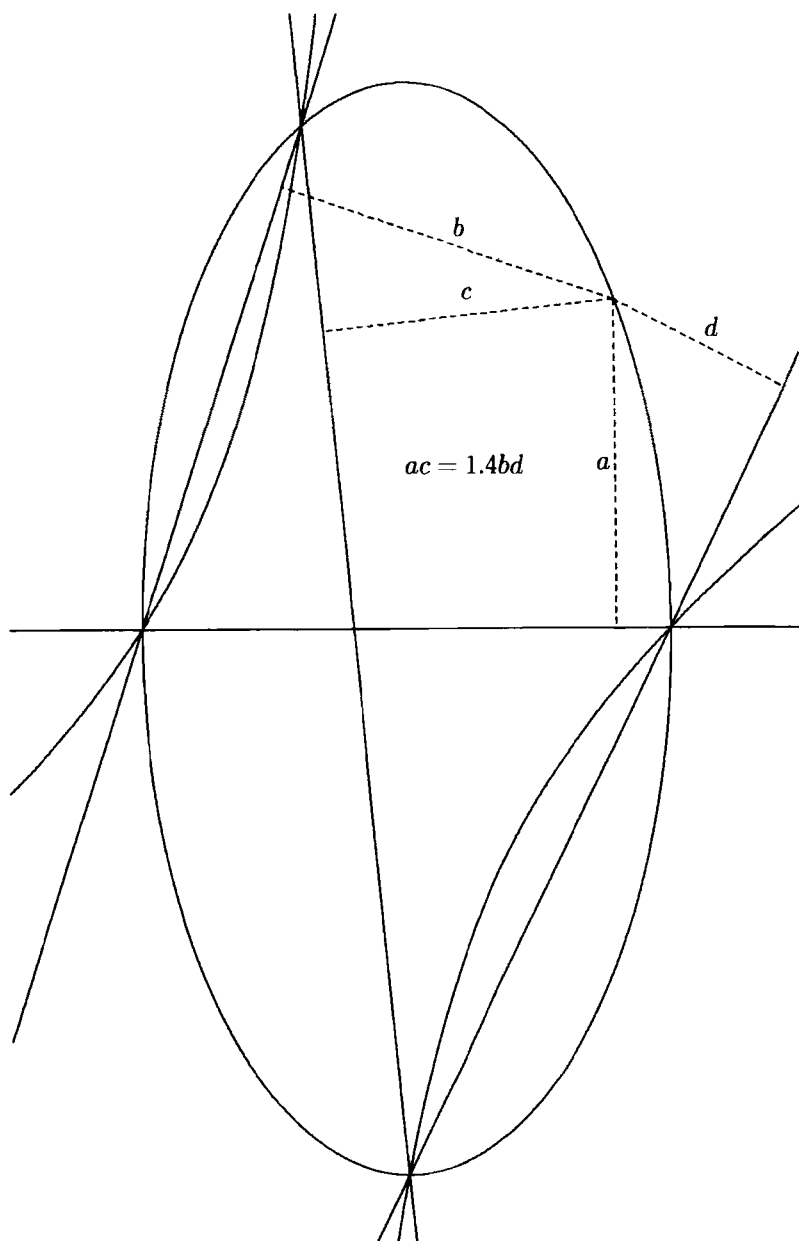


FIGURE 23. The four-line locus. If a point moves so that the product of its distances to two lines bears a constant ratio to the product of its distances to two other lines, it must move in a conic. In this illustration, two conics satisfy the condition: one an ellipse, the other a hyperbola.

10.9. In the Pythagorean tradition there were two kinds of mathematical activity. One kind, represented by the attempt to extend the theory of the transformation of polygons to circles and solid figures, is an attempt to discover new facts and enlarge the sphere of mathematics—to generalize. The other, represented by the discovery

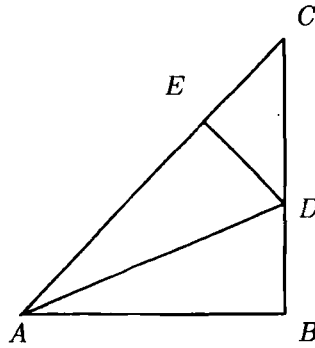


FIGURE 24. Diagonal and side of a square.

of incommensurables, is an attempt to bring into sharper focus the theorems already proved and to test the underlying assumptions of a theory—to rigorize. Are these kinds of activity complementary, opposed, or simply unrelated to each other?

10.10. Hippocrates' quadrature of a lune used the fact that the areas of circles are proportional to the squares on their radii. Could Hippocrates have known this fact? Could he have proved it?

10.11. Plato apparently refers to the famous 3–4–5 right triangle in the *Republic*, 546c. Proclus alludes to this passage in a discussion of right triangles with commensurable sides. We can formulate the recipes that Proclus attributes to Pythagoras and Plato respectively as

$$(2n+1)^2 + (2n^2+2n)^2 = (2n^2+2n+1)^2$$

and

$$(2n)^2 + (n^2-1)^2 = (n^2+1)^2.$$

Considering that Euclid's treatise is regarded as a compendium of Pythagorean mathematics, why is this topic not discussed? In which book of the *Elements* would it belong?

10.12. Proposition 14 of Book 2 of Euclid shows how to construct a square equal in area to a rectangle. Since this construction is logically equivalent to constructing the mean proportional between two line segments, why does Euclid wait until Book 6, Proposition 13 to give the construction of the mean proportional?

10.13. Show that the problem of squaring the circle is equivalent to the problem of squaring one segment of a circle when the central angle subtended by the segment is known. (Knowing a central angle means having two line segments whose ratio is the same as the ratio of the angle to a full revolution.)

10.14. Referring to Fig. 18, show that all the right triangles in the figure formed by connecting B' with C , C' with K , and K' with L are similar. Write down a string of equal ratios (of their legs). Then add all the numerators and denominators to deduce the equation

$$(BB' + CC' + \cdots + KK' + LM) : AM = A'B : BA.$$

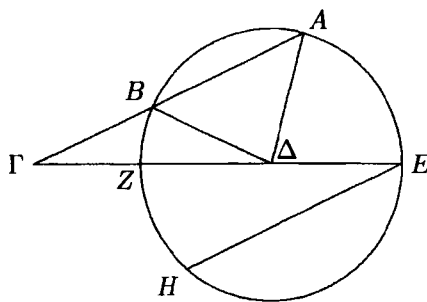


FIGURE 25. Archimedes' trisection of an angle: $\angle A\Gamma\Delta = \frac{1}{3}\angle A\Delta E$.

10.15. Show that Archimedes' result on the relative volumes of the sphere, cylinder, and cone can be obtained by considering the cylinder, sphere and double-napped cone formed by revolving a circle inscribed in a square about a midline of the square, the cone being generated by the diagonals of the square. In this case the area of a circular section of the cone plus the area of the same section of the sphere equals the area of the section of the cylinder since the three radii form the sides of a right triangle. The radius of a section of the sphere cuts off a segment of the axis of rotation from the center equal to the radius of the section of the cone, since the vertex angle of the cone is a right angle. These two segments form the legs of a right triangle whose hypotenuse is a radius of the sphere, which is equal to the radius of the section of the cylinder.

10.16. A minor work attributed to Archimedes called the *Book of Lemmas* contains an angle trisection. In Fig. 25 we are given an acute angle $\angle A\Delta E$, whose trisection is required. We draw a circle of any radius r about Δ , the vertex of the angle. Then, using a straightedge, we mark off on it two points P and Q separated by the distance r . Setting the straightedge down so that P is at point Γ on the extension of the diameter $E\Delta Z$, Q is at point B on the circle, and the point A is also on the edge of the straightedge, we draw the line $A\Gamma$. By drawing EH parallel to $A\Gamma$, we get $\angle A\Gamma E = \angle \Gamma E H$. By joining ΔB , we obtain the isosceles triangle $\Gamma B\Delta$. Now since $\angle B\Delta Z$ is a central angle on the arc \widehat{BZ} and is equal to $\angle B\Gamma\Delta$, which is equal to $\angle ZEH$, which is inscribed in the arc \widehat{ZH} , it follows that $\widehat{ZH} = 2\widehat{BZ}$. Since the arcs \widehat{AE} and \widehat{BH} are equal (being cut off by parallel chords), we now get $\widehat{AE} = \widehat{BH} = 3\widehat{BZ}$. Therefore, $\angle A\Gamma E = \angle B\Delta Z = \frac{1}{3}\angle A\Delta E$.

Why is this construction *not* a straightedge-and-compass trisection of the angle, which is known to be impossible? How does it compare with the *neûsis* trisection shown above? Show how to obtain this same result more simply by erasing everything in the figure below the diameter of the circle.

10.17. Show that the problem of increasing the size of a sphere by half is equivalent to the problem of two mean proportionals (doubling the cube).

10.18. A circle can be regarded as a special case of an ellipse. What is the *latus rectum* of a circle?

10.19. When the equation $y^2 = Cx - kx^2$ is converted to the standard form

$$\frac{(x-h)^2}{a^2} + \frac{y^2}{b^2} = 1,$$

what are the quantities h , a , and b in terms of C and k ?

10.20. Show from Apollonius' definition of the foci that the product of the distances from each focus to the ends of the major axis of an ellipse equals the square on half of the minor axis.

10.21. We have seen that the three- and four-line locus problems have conic sections as their solutions. State and solve the two-line locus problem. You may use modern analytic geometry and assume that the two lines are the x axis and the line $y = ax$. The locus is the set of points whose distances to these two lines have a given ratio. What curve is this?

10.22. Show that the apparent generality of Apollonius' statement of the three-line locus problem, in which arbitrary angles can be prescribed at which lines are drawn from the locus to the fixed lines, is illusory. (To do this, show that the ratio of a line from a point P to line l making a fixed angle θ with the line l bears a constant ratio to the line segment from P perpendicular to l . Hence if the problem is solved for all ratios in the special case when lines are drawn from the locus perpendicular to the given lines, then it is solved for all ratios in any case.)

10.23. Show that the line segment from a point $P = (x, y)$ to a line $ax + by = c$ making angle θ with the line has length

$$\frac{|ax + by - c|}{\sqrt{a^2 + b^2} \sin \theta}.$$

Use this expression and three given lines $l_i : a_i x + b_i y = c_i$, $i = 1, 2, 3$, to formulate the three-line locus problem analytically as a quadratic equation in two variables by setting the square of the distance from (x, y) to line l_1 equal to a constant multiple of the product of the distances to l_2 and l_3 . Show that the locus passes through the intersection of the line l_1 with l_2 and l_3 , but not through the intersection of l_2 with l_3 . Also show that its tangent line where it intersects l_i is l_i itself, $i = 2, 3$.

10.24. One reason for doubting Cavalieri's principle is that it breaks down in one dimension. Consider, for instance, that every section of a right triangle parallel to one of its legs meets the other leg and the hypotenuse in congruent figures (a single point in each case). Yet the other leg and the hypotenuse are obviously of different lengths. Is there a way of redefining "sections" for one-dimensional figures so that Cavalieri's principle can be retained? If you could do this, would your confidence in the validity of the principle be restored?

10.25. We know that interest in conic sections *arose* because of their application to the problem of two mean proportionals (doubling the cube). Why do you think interest in them was *sustained* to the extent that caused Euclid, Aristaeus, and Apollonius to write treatises developing their properties in such detail?

10.26. Pappus' history of the conics implies that people knew that the ellipse, for example, could be obtained by cutting a right-angled cone with a plane. Can *every* ellipse be obtained by cutting a right-angled cone with a plane? Prove that it can, by showing that any a and b whatsoever in Eq. 2 can be obtained as the section of the

right-angled cone whose equation is $y^2 = zx$ by the plane $x = 2a - (a^2z/b^2)$. Then show that by taking $a = eu/(1 - e^2)$, $b = a\sqrt{1 - e^2}$, $x = w$, $y = v$, where $e = h/w$, you get Eq. 1. [*Hint*: Recall that e is constant in a given conic section. Also, observe that $0 < e < 1$ for a section of an acute-angle cone, since $h = w \tan(\theta/2)$, where θ is the vertex angle of the cone.]

10.27. As we have seen, Apollonius was aware of the string property of ellipses, yet he did not mention that this property could be used to draw an ellipse. Do you think that he did not *notice* this fact, or did he omit to mention it because he considered it unimportant?

10.28. Prove Proposition 54 of Book 3 of Apollonius' *Conics* in the special case in which the conic is a circle and the point Θ is at the opposite end of the diameter from B (Fig. 22).

CHAPTER 11

Post-Euclidean Geometry

A certain dullness came over Greek geometry from the beginning of the second century BCE. The preceding century had seen the beginning of Roman expansion, whose early stage took the life of the aged Archimedes. Julius Caesar (100–44 BCE), who did more than anyone before him to turn Rome from a republic into an empire, took an army to Egypt to fight his rival Pompey and incidentally help Cleopatra, the last of the heirs of Ptolemy Soter, defeat her brother in a civil war. In pursuing his aim he sent fire ships into the harbor of Alexandria to set it ablaze. Although he himself naturally says nothing about any destruction of the city, later writers, such as Plutarch in his *Life of Caesar* and Gellius in his *Attic Nights*, say that the fire damaged the Library. Gellius claims that 700,000 books were destroyed. After Caesar's heir Octavian defeated Mark Antony and Cleopatra in 31 BCE, Egypt became a province of the Roman Empire. Whether because of this disruption or from limitations inherent in the Pythagorean philosophy, the level of brilliant achievements of Euclid, Archimedes, and Apollonius was not sustained. Nevertheless, geometry did not die out entirely, and some of the later commentators are well worth reading. Very little new geometry was written in Greek after the sixth century, however. From the ninth century to the fifteenth the Euclidean tradition in geometry was pursued by Muslim mathematicians. Since these mathematicians were also interested in the philosophy of Aristotle, in their work mathematics once again began to be mixed with philosophy, as it was in the time just before Euclid.

When the Roman Empire was vigorous, all upper-class Romans understood Greek, and many seemed to prefer it to Latin. The Emperor Marcus Aurelius, for example, who ruled from 161 to 180, wrote his meditations in Greek. After the Emperor Diocletian (284–305) split the empire into eastern and western halves to make it governable and the eastern Emperor Constantine (307–337), who proclaimed Christianity the official religion of the Empire, moved his capital to Constantinople, knowledge of Greek began to decline in the western part of the empire. Many books were translated into Latin, or replacements for them were written in Latin.

The repeated ravaging of Italy by invaders from the north caused an irreversible decline in scholarship there. In the east, which fared somewhat better, scholarship continued for another thousand years, until the Turkish conquest of Constantinople in 1453. The eastern Emperor Justinian (525–565) managed to reassert his rule over part of Italy, but this project proved too expensive to sustain, and Italy was soon once again beyond the control of the Emperor. For several centuries before the reign of Justinian an entirely new civilization based on Christianity had been replacing the ancient Greco-Roman world, symbolically marked by the Justinian's closing of the pagan Academy at Athens in 529.

1. Hellenistic geometry

Although the Euclidean restrictions set limits to the growth of geometry, there remained people who attempted to push the limits beyond the achievements of Archimedes and Apollonius, and they produced some good work over the next few centuries. We shall look at just a few of them.

1.1. Zenodorus. The astronomer Zenodorus lived in Athens in the century following Apollonius. Although his exact dates are not known, he is mentioned by Diocles in his book *On Burning Mirrors* and by Theon of Smyrna. According to Theon, Zenodorus wrote *On Isoperimetric Figures*, in which he proved four theorems: (1) If two regular polygons have the same perimeter, the one with the larger number of sides encloses the larger area; (2) a circle encloses a larger area than any regular polygon whose perimeter equals its circumference; (3) of all polygons with a given number of sides and perimeter, the regular polygon is the largest; (4) of all closed surfaces with a given area, the sphere encloses the largest volume. With the machinery inherited from Euclidean geometry, Zenodorus could not have hoped for any result more general than these. Let us examine his proof of the first two, as reported by Theon.

Referring to Fig. 1, let $AB\Gamma$ and ΔEZ be two regular polygons having the same perimeter, with $AB\Gamma$ having more sides than ΔEZ . Let H and Θ be the centers of these polygons, and draw the lines from the centers to two adjacent vertices and their midpoints, getting triangles $B\Gamma H$ and $EZ\Theta$ and the perpendicular bisectors of their bases HK and $\Theta\Lambda$. Then, since the two polygons have the same perimeter but $AB\Gamma$ has more sides, BK is shorter than $E\Lambda$. Mark off M on $E\Lambda$ so that $M\Lambda = BK$. Then if P is the common perimeter, we have $E\Lambda : P :: \angle E\Theta\Lambda : 4 \text{ right angles}$ and $P : BK :: 4 \text{ right angles} : \angle BHK$. By composition, then $E\Lambda : BK :: \angle E\Theta\Lambda : \angle BHK$, and therefore $E\Lambda : M\Lambda :: \angle E\Theta\Lambda : \angle BHK$. But, Zenodorus claimed, the ratio $E\Lambda : M\Lambda$ is *larger* than the ratio $\angle E\Theta\Lambda : \angle M\Theta\Lambda$, asking to postpone the proof until later. Granting that lemma, he said, the ratio $\angle E\Theta\Lambda : \angle BHK$ will be larger than the ratio $\angle E\Theta\Lambda : \angle M\Theta\Lambda$, and therefore $\angle BHK$ is smaller than $\angle M\Theta\Lambda$. It then follows that the complementary angles $\angle HBK$ and $\angle \Theta M\Lambda$ satisfy the reverse inequality. Hence, copying $\angle HBK$ at M so that one side is along $M\Lambda$, we find that the other side intersects the extension of $\Lambda\Theta$ at a point N beyond Θ . Then, since triangles BHK and MNA are congruent by angle-side-angle, it follows that $HK = NA > \Theta\Lambda$. But it is obvious that the areas of the two polygons are $\frac{1}{2}HK \cdot P$ and $\frac{1}{2}\Theta\Lambda \cdot P$, and therefore $AB\Gamma$ is the larger of the two.

The proof that the ratio $E\Lambda : M\Lambda$ is larger than the ratio $\angle E\Theta\Lambda : \angle M\Theta\Lambda$ was given by Euclid in his *Optics*, Proposition 8. But Theon does not cite Euclid in his quotation of Zenodorus. He gives the proof himself, implying that Zenodorus did likewise. The proof is shown on the top right in Fig. 1, where the circular arc ΞMN has been drawn through M with Θ as center. Since the ratio $\triangle E\Theta M : \text{sector } N\Theta M$ is larger than the ratio $\triangle M\Theta\Lambda : \text{sector } M\Theta\Xi$ (the first triangle is larger than its sector, the second is smaller), it follows, interchanging means, that $E\Theta M : M\Theta\Lambda > \text{sector } N\Theta M : \text{sector } M\Theta\Xi$. But $E\Theta M : M\Theta\Lambda :: EM : M\Lambda$, since the two triangles have the same altitude measured from the base line $EM\Lambda$. And $\text{sector } N\Theta M : \text{sector } M\Theta\Xi :: \angle E\Theta M : \angle M\Theta\Lambda$. Therefore, $EM : M\Lambda$ is larger than the ratio $\angle E\Theta M : \angle M\Theta\Lambda$, and it then follows that $E\Lambda : M\Lambda$ is larger than $\angle E\Theta\Lambda : \angle M\Theta\Lambda$.

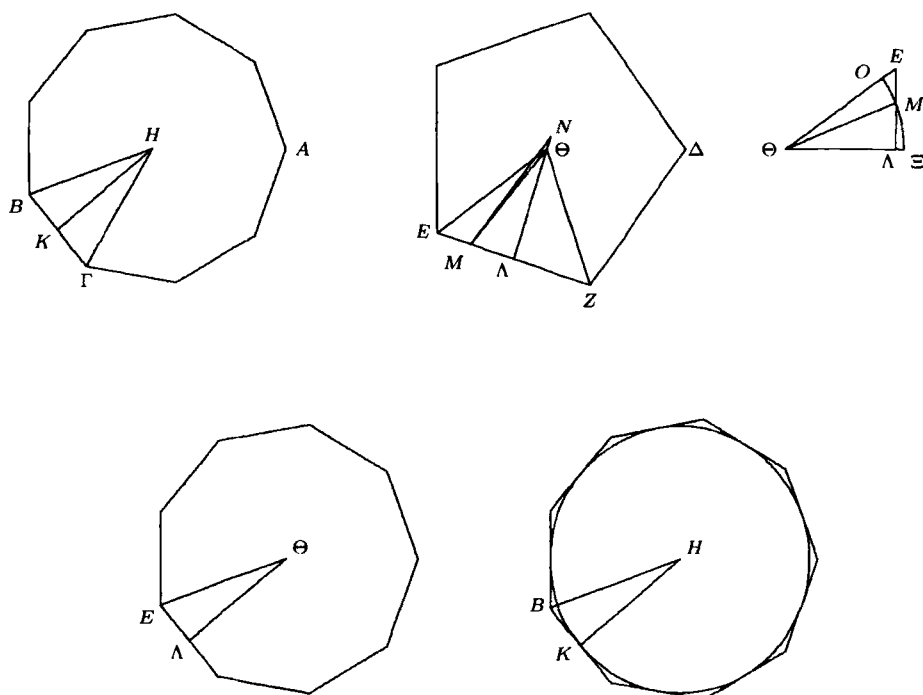


FIGURE 1. Two theorems of Zenodorus. Top: When two regular polygons have the same perimeter, the one with the larger number of sides is larger. Bottom: A circle is larger than a regular polygon whose perimeter equals the circumference of the circle.

Zenodorus' proof that a circle is larger than a regular polygon whose perimeter equals the circumference of the circle is shown at the bottom of Fig. 1. Given such a polygon and circle, circumscribe a similar polygon around the circle. Since this polygon is "convex on the outside," as Archimedes said in his treatise on the sphere and cylinder, it can be assumed longer than the circumference. (Both Archimedes and Zenodorus recognized that this was an assumption that they could not prove; Zenodorus cited Archimedes as having assumed this result.) That means the circumscribed polygon is larger than the original polygon since it has a larger perimeter. But then by similarity, HK is larger than $\Theta\Lambda$. Since the area of the circle equals half of the rectangle whose sides are its circumference and HK , while the area of the polygon is half of the rectangle whose sides are its perimeter and $\Theta\Lambda$, it follows that the circle is larger.

1.2. The parallel postulate. We saw in the Chapter 10 that there was a debate about the theory of parallel lines in Plato's Academy, as we infer from the writing of Aristotle. This debate was not ended by Euclid's decision to include a parallel postulate explicitly in the *Elements*. This foundational issue was discussed at length

by the Stoic philosopher Geminus, whose dates are a subject of disagreement among experts, but who probably lived sometime between 50 BCE and 50 CE. Geminus wrote an encyclopedic work on mathematics, which has been entirely lost, except for certain passages quoted by Proclus, Eutocius, and others. Proclus said that the parallel postulate should be completely written out of the list of postulates, since it is really a theorem. The asymptotes of hyperbolas provided the model on which he reasoned that converging is not the same thing as intersecting. But still he thought that such behavior was impossible for straight lines. He claimed that a line that intersected one of two parallel lines must intersect the other,¹ and he reports a proof of Geminus that assumes in many places that certain lines drawn will intersect, not realizing that by doing so he was already assuming the parallel postulate.

Proclus also reports an attempt by Ptolemy to prove the postulate by arguing that a pair of lines could not be parallel on one side of a transversal “rather than” on the other side. (Proclus did not approve of this argument.) But of course the assumption that parallelism is two-sided is one of the properties of Euclidean geometry that does not extend to hyperbolic geometry. These early attempts to prove the parallel postulate began the process of unearthing more and more plausible alternatives to the postulate, but of course did not lead to a proof of it.

1.3. Heron. We have noted already the limitations of the Euclidean approach to geometry, the chief one being that lengths are simply represented as lines, not numbers. After Apollonius, however, the metric aspects of geometry began to resurface in the work of later writers. One of these writers was Heron (ca. 10–ca. 75), who wrote on mechanics; he probably lived in Alexandria. Pappus discusses his work at some length in Book 8 of his *Synagōgē*. Heron’s geometry is much more concerned with measurement than was the geometry of Euclid. The change of interest in the direction of measurement and numerical procedures signaled by his *Metrika* is shown vividly by his repeated use (130 times, to be exact) of the word *area* (*embadón*), a word never once used by Euclid, Archimedes, or Apollonius.² There is a difference in point of view between saying that two plane figures are equal and saying that they have the same area. The first statement is geometrical and is the stronger of the two. The second is purely numerical and does not necessarily imply the first. Heron discusses ways of finding the areas of triangles from their sides. After giving several examples of triangles that are either integer-sided right triangles or can be decomposed into such triangles by an altitude, such as the triangle with sides of length 13, 14, 15, which is divided into a 5–12–13 triangle and a 9–12–15 triangle by the altitude to the side of length 14, he gives “a direct method by which the area of a triangle can be found without first finding its altitude.” He

¹ This assertion is an *assumption* equivalent to the parallel postulate and obviously equivalent to the form of the postulate commonly used nowadays, known as Playfair’s axiom: *Through a given point not on a line, only one parallel can be drawn to the line.*

² Reporting (in his commentary on Ptolemy’s *Almagest*) on Archimedes’ measurement of the circle, however, Theon of Alexandria did use this word to describe what Archimedes did; but that usage was anachronistic. In his work on the sphere, for example, Archimedes referred to its *surface* (*epipháneia*), not its *area*. On the other hand, Dijksterhuis (1956, pp. 412–413) reports the Arabic mathematician al-Biruni as having said that “Heron’s formula” is really due to Archimedes. Considering the contrast in style between the proof and the applications, it does appear plausible that Heron learned the proof from Archimedes. Heath (1921, Vol. 2, p. 322) endorses this assertion unequivocally.

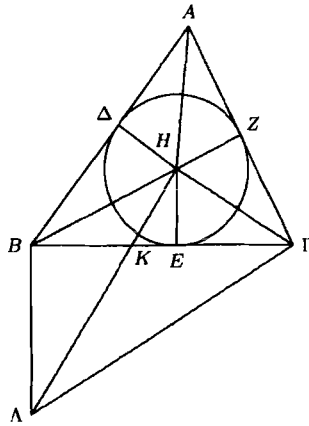


FIGURE 2. Heron's proof of his direct method of computing the area of a triangle.

gave as an example a triangle whose sides were 7, 8, and 9 units. His prescription was: Add 9 and 8 and 7, getting 24. Take half of this, getting 12. Subtract 7 units from this, leaving 5. Then subtract 8 from 12, leaving 4. Finally, subtract 9, leaving 3. Multiply 12 by 5, getting 60. Multiply this by 4, getting 240. Multiply this by 3, getting 720. Take the square root of this, and that will be the area of the triangle. He went on to explain that since 720 is not a square, it will be necessary to approximate, starting from the nearest square number, 729.

This result seems anomalous in Greek geometry, since Heron is talking about multiplying an area by an area. That is probably why he emphasizes that his results are numerical rather than geometric. An examination of his proof of the formula shows that he need not have multiplied two areas together. He must have made a deliberate choice to express himself this way. His proof is based on Fig. 2, in which one superfluous line has been omitted to streamline it. In the following proof, some rewording has been done to accommodate this minor modification of the figure.

The lines ΛB and ΛH are perpendicular respectively to $B\Gamma$ and $H\Gamma$. The proof follows easily once it is shown that the quadrilateral $\Lambda B H \Gamma$ is cyclic, that is, can be inscribed in a circle. In fact, if Σ denotes the semiperimeter, then

$$\begin{aligned} \Sigma^2 : \Sigma \cdot (\Sigma - B\Gamma) :: \Sigma : (\Sigma - B\Gamma) :: (\Sigma - A\Gamma) : KE :: (\Sigma - A\Gamma) \cdot \Gamma E : KE \cdot \Gamma E \\ = (\Sigma - A\Gamma) \cdot (\Sigma - AB) : EH^2. \end{aligned}$$

Here $KE \cdot \Gamma E = EH^2$ because EH is the altitude to the base of the right triangle HKT .

Heron *could have* stated the result in Euclidean language if he had wanted to. If he were to regard each term in the proportion

$$\Sigma^2 : \Sigma \cdot (\Sigma - B\Gamma) :: (\Sigma - A\Gamma) \cdot (\Sigma - AB) : EH^2$$

as an area and take the sides of squares equal to them, he would have four squares in proportion, of sides Σ , α , β , EH , where α is the mean proportional between Σ and $\Sigma - B\Gamma$ and β the mean proportional between $\Sigma - A\Gamma$ and $\Sigma - AB$. It would need to be proved that if four squares are in proportion, then their sides are also in proportion; however, that fact follows immediately from the Eudoxan theory of

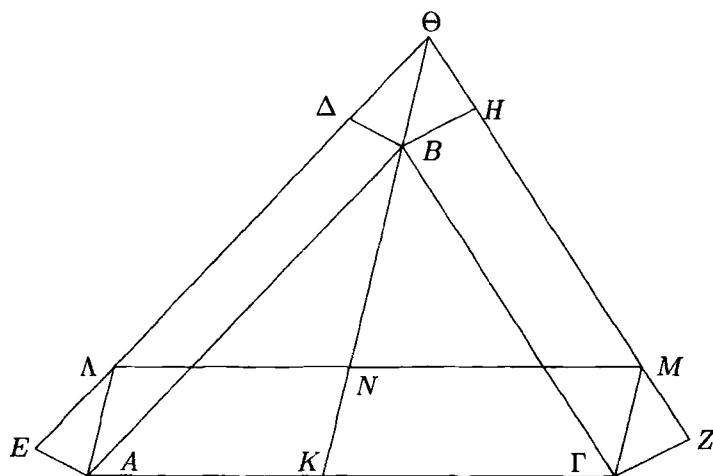


FIGURE 3. Pappus' generalization of the Pythagorean theorem.

proportion (see Problem 11.3). Working with the sides of the squares, it would then be legitimate to multiply means and extremes—that is, to form rectangles on the sides—since the appropriate theorems were proved in Book 6 of Euclid. He could have said that the triangle $AB\Gamma$ equals the rectangle on Σ and EH , which in turn equals the rectangle on α and β . The assertion that the triangle $AB\Gamma$ is the rectangle on α and β is precisely Heron's theorem. What he has done up to this point would not have offended a logical Euclidean purist. Why did he not finish the proof in this way?

The most likely explanation is that the proof came from Archimedes, as many scholars believe, and that Heron was aiming at numerical results. Another possible explanation is that our reconstruction of what Heron *could have* done lacks the symmetry of the process described by Heron, since α and β do not contain the sides in symmetric form. Whatever the reason, his summing up of the argument leaves no doubt that he was willing to accept the product of two areas as a product of numbers.

1.4. Pappus. Book 4 of Pappus' *Synagōgē* contains a famous generalization of the Pythagorean theorem: Given a triangle $AB\Gamma$ and any parallelograms $B\Gamma ZH$ and $AB\Delta E$ constructed on two sides, it is possible to construct (with straightedge and compass) a parallelogram $A\Gamma M\Lambda$ on the third side equal in area to the sum of the other two (see Fig. 3).

The isoperimetric problem. In Book 5 Pappus states almost verbatim the argument that Thcon of Alexandria, quoting Zenodorus, gave for the proof of the isoperimetric inequality. Pappus embroiders the theorem with a beautiful literary device, however. He speaks poetically of the divine mission of the bees to bring from heaven the wonderful nectar known as honey and says that in keeping with this mission they must make their honeycombs without any cracks through which honey could be lost. Being endowed with a divine sense of symmetry as well, the bees had to choose among the regular shapes that could fulfill this condition, that is, triangles, squares, and hexagons. They chose the hexagon because a hexagonal

prism required the least material to enclose a given volume, out of all the possible prisms whose base would tile the plane.³

Analysis, locus problems, and Pappus' theorem. Book 7 of the *Synagōgē* is a treasure trove of fascinating information about Greek geometry for several reasons. First, Pappus describes the kinds of techniques used to carry on the research that was current at the time. He lists a number of books of this *analysis* and tells who wrote them and what their contents were, in general terms, thereby providing valuable historical information. What he means by analysis, as opposed to synthesis, is a kind of algebraic reasoning in geometry. As he puts it, when a construction is to be made or a relation is to be proved, one imagines the problem to have been solved and then deduces consequences connecting the result with known principles, after which the process is reversed and a proof can be synthesized. This process amounts to thinking about objects not yet determined in terms of properties that they must have; when applied to numbers, that process is algebra.

Second, Book 7 also contains a general discussion of locus problems, such as we have already encountered in Apollonius' *Conics*. This discussion exerted a strong influence on the development of geometry in seventeenth-century France.

Proposition 81 of Euclid's *Data*, discussed above, inspired Pappus to create a very general proposition about plane loci. Referring to the points of intersection of a set of lines, he writes:

To combine these discoveries in a single proposition, we have written the following. *If three points are fixed on one line... and all the others except one are confined to given lines, then that last one is also confined to a given line.* This is asserted only for four lines, no more than two of which intersect in the same point. It is not known whether this assertion is true for every number.

Pappus could not have known that he had provided the essential principle by which a famous theorem of projective geometry known as Desargues' theorem (see Section 2 of Chapter 12) was to be proved 1400 years later. Desargues certainly knew the work of Pappus, but may not have made the connection with this theorem. The connection was pointed out by van der Waerden (1963, p. 287).

Pappus discusses the three- and four-line locus for which the mathematical machinery is found in Book 3 of Apollonius' *Conics*. For these cases the locus is always one of the three conic sections. Pappus mentions that the two-line locus is a planar problem; that is, the solution is a line or circle. He says that a point satisfying the conditions of the locus to five or six lines is confined to a definite curve (a curve "given in position" as the Greeks said), but that this curve is "not yet familiar, and is merely called a curve." The curve is defined by the condition that the rectangular parallelepiped spanned by the lines drawn from a point to three fixed lines bears a fixed ratio to the corresponding parallelepiped spanned by the lines drawn to three other fixed lines. In our terms, this locus is a cubic curve.

³ If one is looking for mathematical explanations of this shape, it would be simpler to start with the assumption that the body of a bee is approximately a cylinder, so that the cells should be approximately cylinders. Now one cylinder can be tightly packed with six adjacent cylinders of the same size. If the cylinders are flexible and there is pressure on them, they will flatten into hexagonal prisms.

Third, in connection with the extension of these locus problems, Pappus considers the locus to more than six lines and says that a point satisfying the corresponding conditions is confined to a definite curve. This step was important, since it proposed the possibility that a curve could be determined by certain conditions without being explicitly constructible. Moreover, it forced Pappus to go beyond the usual geometric interpretation of products of lines as rectangles, thus pushing the same boundary that Heron had gone through. Noting that "nothing is subtended by more than three dimensions," he continues:

It is true that some of our recent predecessors have agreed among themselves to interpret such things, but they have not made a meaningful clear definition in saying that what is subtended by certain things is multiplied by the square on one line or the rectangle on others. But these things can be stated and proved using the composition of ratios.

It appears that Pappus was on the very threshold of the creation of the modern concept of a real number as a ratio of lines. Why did he not cross that threshold? The main reason was probably the cumbersome Euclidean definition of a composite ratio, discussed in Chapter 10. But there was a further reason: he wasn't interested in foundational questions. He made no attempt to prove or justify the parallel postulate, for example. And that brings us to the fourth attraction of Book 7. In that book Pappus investigated some very interesting problems, which he preferred to foundational questions. After concluding his discussion of the locus problems, he implies that he is merely reporting what other people, who are interested in them, have claimed. "But," he says,

after proving results that are much stronger and promise many applications... to show that I do not come boasting and empty-handed... I offer my readers the following: *The ratio of rotated bodies is the composite of the ratio of the areas rotated and the ratio of straight lines drawn similarly [at the same angle] from their centers of gravity to the axes of rotation. And the ratio of incompletely rotated bodies is the composite of the ratio of the areas rotated and the ratio of the arcs described by their centers of gravity.*

Pappus does not say how he discovered these results, nor does he give the proof. The proof would have been fairly easy, given that he had read Archimedes' *Quadrature of the Parabola*, in which the method of exhaustion is used. For the first theorem it would have been sufficient to compute the volume generated by revolving a right triangle with one leg parallel to the axis of rotation, and in that case the volume could be computed by subtracting the volume of a cylinder from the volume of a frustum of a cone. If the theorem is true for two nonoverlapping areas, it is easily seen to be true for the union of those areas. Pappus could then have applied the method of exhaustion to get the general result. The second result is an immediate application of the Eudoxan theory of proportion, since the volume generated is obviously in direct proportion to the angle of rotation, as are the arcs traversed by individual points. The modern theorem that is called Pappus' theorem asserts that the volume of a solid of revolution is equal to the product of

the area rotated and the distance traversed by its center of gravity (which is 2π times the length of the line from the center of gravity to the axis of rotation). In the modern form this theorem was first stated in 1609 by the Swiss astronomer-mathematician Paul Guldin (1577–1643), a Jesuit priest, and published between 1635 and 1640 in the second volume of his four-volume work *Centrobaryca seu de centro gravitatis trium specierum quantitatis continuæ* (*The Barycenter, or on the Center of Gravity of the Three Kinds of Continuous Magnitude*). Guldin had apparently not read Pappus and made the discovery independently. He did not prove the result, and the first proof is due to Bonaventura Cavalieri (1598–1647).

2. Roman geometry

In the Roman Empire geometry found applications in mapmaking. The way back from the abstractness of Euclidean geometry was led by Heron, Ptolemy, and other geometers who lived during the early Empire. We have already mentioned Ptolemy's *Almagest*, which was an elegant arithmetization of some basic Euclidean geometry applied to astronomy. In it, concrete computations using the table of chords are combined with rigorous geometric demonstration of the relations involved. But Ptolemy studied the Earth as well as the sky, and his contribution to geography is also a large one, and also very geometric.

Ptolemy was one of the first scholars to look at the problem of representing large portions of the Earth's surface on a flat map. His data, understandably very inaccurate from the modern point of view, came from his predecessors, including the astronomers Eratosthenes (276–194) and Hipparchus (190–120) and the geographers Strabo (ca. 64 BCE–24 CE) and Marinus of Tyre (70–130), whom he followed in using the now-familiar lines of latitude and longitude. These lines have the advantage of being perpendicular to one another, but the disadvantage that the parallels of latitude are of different sizes. Hence a degree of longitude stands for different distances at different latitudes.

Ptolemy assigned latitudes to the inhabited spots that he knew about by computing the length of sunlight on the longest day of the year. This computational procedure is described in Book 2, Chapter 6 of the *Almagest*, where Ptolemy describes the latitudes at which the longest day lasts $12\frac{1}{4}$ hours, $12\frac{1}{2}$ hours, and so on up to 18 hours, then at half-hour intervals up to 20 hours, and finally at 1-hour intervals up to 24. Although he knew theoretically what the Arctic Circle is, he didn't know of anyone living north of it, and took the northernmost location on the maps in his *Geography* to be Thoûlē, described by the historian Polybius around 150 BCE as an island six days sail north of Britain that had been discovered by the merchant-explorer Pytheas (380–310) of Masillia (Marseille) some two centuries earlier.⁴ It has been suggested that Thoûlē is the Shetland Islands (part of Scotland since 1471), located between 60° and 61° north; that is just a few degrees south of the Arctic Circle, which is at $66^\circ 30'$. It is also sometimes said to be Iceland, which is on the Arctic Circle, but west of Britain rather than north. Whatever it was, Ptolemy assigned it a latitude of 63° , although he said in the *Almagest* that some "Scythians" (Scandinavians and Slavs) lived still farther north at $64\frac{1}{2}^\circ$. Ptolemy did know of people living south of the equator and took account of places as far south as Agisymba (Ethiopia) and the promontory of Prasum (perhaps Cabo Delgado in Mozambique, which is 14° south). Ptolemy placed it $12^\circ 30'$ south of

⁴ The Latin idiom *ultima Thule* means roughly *the last extremity*.

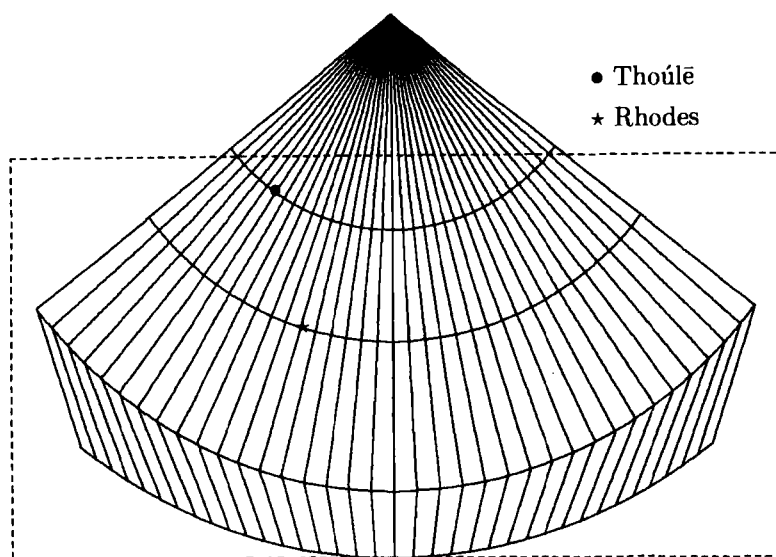


FIGURE 4. Ptolemy's first method of mapping.

the Equator. The extreme southern limit of his map was the circle $16^{\circ} 25'$ south of the equator, which he called "anti-Meróē," since Meróē was $16^{\circ} 25'$ north.

Since he knew only the geography of what is now Europe, Africa, and Asia, he did not need 360° of longitude. He took his westernmost point to be the Blessed Islands (possibly the Canary Islands, at 17° west). That was his prime meridian, and he measured longitude out to 180° eastward from there, to the Sêres⁵ and the Chinese (Sínai) and "Kattígara." According to Dilke (1985, p. 81), "Kattígara" may refer to Hanoi. Actually, the east-west span from the Canary Islands to Shanghai (about 123° east) is only 140° of longitude. Ptolemy's inaccuracy is due partly to unreliable reports of distances over trade routes and partly to his decision to accept 500 stades as the length of a degree of latitude when the true distance is about 600 stades.⁶ We are not concerned with geography, however, only with its mathematical aspects.

The problem Ptolemy faced was to draw a flat map of the Earth's surface spanning 180° of longitude and about 80° degrees of latitude, from $16^{\circ} 25'$ south to 63° north. Ptolemy described three methods of doing this, the first of which we shall now discuss. The latitude and longitude coordinates of the inhabited world (*oikuménē*) known to Ptolemy represent a rectangle whose width is $\frac{5}{9}$ of its length. But Ptolemy did not like to represent parallels of latitude as straight lines; he preferred to draw them as arcs of concentric circles while keeping the meridians

⁵ The Sêres were a Hindu people known to the Greeks from the silk trade.

⁶ It has become a commonplace that Christopher Columbus, relying on Ptolemy's geography, expected to reach the Orient at a distance that would have placed him in the middle of the Pacific Ocean had North America not been in the way. If he believed Ptolemy, he would have thought it about 180° of longitude, which at a latitude of 40° would have been about 138 great-circle degrees. But he thought a degree was 500 stades (92 km), and hence that the distance to Japan was about 12,700 km. Since a degree is actually 600 stades (110 km), the journey would have been more than 15,000 km. But the latitude of Japan is slightly south of the latitude of Spain.

of longitude as straight lines emanating from the common center, representing the north pole. Thus, his plan is to map this portion of the Earth into the portion of a sector of a disk bounded by two radii and two concentric circles. In terms of Fig. 4, his first problem is to decide which radii and which circles are to form these boundaries. Ptolemy recognized that it would be impossible in such a map to place all the parallels of latitude at the correct distances from one another and still get their lengths in proportion. He decided to keep his northernmost parallel, through Thoulē, in proportion to the parallel through the equator. That meant these arcs should be in the proportion of about 9:20—to be precise, $\cos(63^\circ)$ in our terms. Since there would be 63 equal divisions between that parallel and the equator, he needed the upper radius x to satisfy $x : (x + 63) :: 9 : 20$. Solving this proportion is not hard, and one finds that $x = 52$, to the nearest integer. The next task was to decide on the angular opening. For this principle he decided, like his predecessor Marinus, to get the parallel of latitude through Rhodes in the correct proportion. Since Rhodes is at 36° latitude, half of the parallel through it amounts to about $\frac{4}{5}$ of the 180° arc of a great circle, which is about 145° . Since the radius of Rhodes must be 79 (27 great-circle degrees more than the radius of Thoulē), he needed the opening angle of the sectors θ to satisfy $\theta : 180^\circ :: 146 : 79\pi$, so that $\theta \approx 106^\circ$. After that, he inserted meridians of longitude every one-third of an hour of longitude (5°) fanning out from the north pole to the equator.

Ptolemy recognized that continuing to draw the parallels of latitude in the same way for points south of the equator would lead to serious distortion, since the circles in the sector continue to increase as the distance south of the north pole increases, while the actual parallels on the Earth begin to decrease at that point. The simplest solution to that problem, he decided, was to let his southernmost parallel at $16^\circ 25'$ south have its actual length, then join the meridians through that parallel by straight lines to the points where they intersect the equator. Once that decision was made, he was ready to draw the map on a rectangular sheet of paper. He gave instructions for how to do that: Begin with a rectangle that is approximately twice as long as it is wide, draw the perpendicular bisector of the horizontal (long) sides, and extend it above the upper edge so that the portion above that edge and the whole bisector are in the ratio $34^\circ : 131^\circ, 25'$. In that way, the 106° arc through Thoulē will begin and end just slightly above the upper edge of the rectangle, while the lowest point of the map will be at the foot of the bisector, being about 80 units below the lowest point on the parallel of Thoulē, as indicated by the dashed line in Fig. 4.

This way of mapping is *not* a conical projection, as it might appear to be, since it preserves north-south distances. It does a tolerably good job of mapping the parts of the world for which Ptolemy had reliable data. One can recognize Europe and the Middle East in the map of Plate 8, constructed around the year 1300 CE to accompany an edition of the *Geography*.

2.1. Roman civil engineering. Dilke (1985, pp. 88–90) describes the use of geometry in Roman civil engineering as follows. The center of a Roman village would be at the intersection of two perpendicular roads, a (usually) north-south road called the *kardo maximus* (literally, the *main hinge*) and an east-west road called *decumanus maximus*, the *main tenth*. Lots were laid out in blocks (*insulae*) called *hundredths* (*centuriæ*), each block being assigned a pair of numbers, telling how many units it was *dextra decumani* (on the right decumanus) or *sinistra decumani*

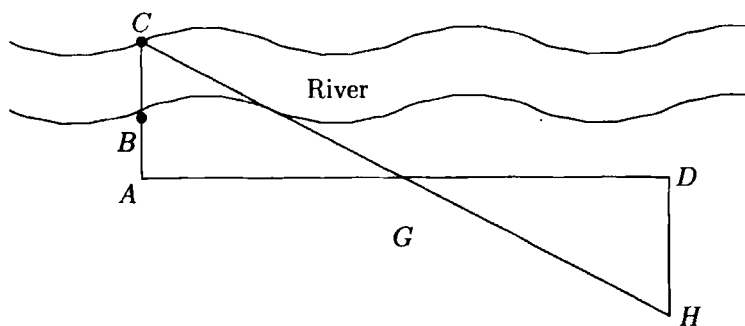


FIGURE 5. Nipsus' method of computing the width of a river.

(on the left decumanus) and how many units it was *ultra kardinem* (on the far kardo) or *citra kardinem* (on the near kardo).⁷

A collection of Roman writings on surveying was collected, translated into German, and published in Berlin in the middle of the nineteenth century. This two-volume work bears the title *Corpus Agrimensorum Romanorum*, the word *agrimensor* (field measurer) being the Latin name for a surveyor. Among the *agrimensores* was one named M. Iunius Nipsus, a second-century surveyor, who, according to Dilke (1985, p. 99), gives the following directions for measuring the width of a river (Fig. 5).

You mark the point *C* on the opposite bank from *B* (a part of the procedure Nipsus neglects to mention until later), continue the straight line *CB* to some convenient point *A*, lay down the crossroads sign at *A*, then move along the direction perpendicular to *AC* until you reach a point *G*, where you erect a pole, then continue on to *D* so that $GD = AG$. You then move away from *D* along the direction perpendicular to *AD* until you see *G* and *C* in a straight line from the point *H*. Since the triangles *AGC* and *DGH* are congruent (by angle-side-angle), it follows that $CB = CA - AB = HD - AB$.

For this procedure to work in practice, it is necessary to have an accessible and level piece of land covering the lines shown as *AD* and *DH*. If the river is large, such a stretch of land may not exist, since the river banks are likely to be hilly. In its neglect of similar triangles, this method seems a large step backward in applied geometry.

3. Medieval geometry

Among the translations of Greek works into Latin mentioned above was a translation of Euclid's *Elements* written by Boethius (ca. 480–524). This work has been lost, although references to it survive.⁸ A "pseudo-Boethius" text of geometry, written some centuries later, has survived. It may have been a standard text during the Middle Ages. There were anonymous treatises on geometry during this time, some attributed to Boethius, usually containing Latin translations of a few of the

⁷ In modern terms these would be *First Avenue East*, *First Avenue West*, *North Main Street*, and *South Main Street*.

⁸ For example, in his *Encyclopedia of Liberal and Literary Studies* the early sixth-century writer Magnus Aurelius Cassiodorus refers to the great Greek mathematicians "of whom Euclid was given to us translated into the Latin language by the same great man Boethius."

early books of Euclid, often drastically edited. The tradition of attributing these works to Boethius continued even in the twelfth century, when translations from Arabic began to appear, as one can see in the booklets of Folkerts (1970, 1971), the second of which compares an anonymous Latin version ascribed to Boethius with the translation (from Arabic) of Athelhard of Bath. In a series of papers (1999, 2000) Zaitsev has argued that the pseudo-Boethius was enlarging the commentaries on Euclid by including material from surveying and geometric astronomy. As a result,

[I]n the writing process geometric concepts were systematically translated into the language of surveying, and the resulting melange of surveying and geometry was used as the basis for discussing the theological–cosmological significance of the discipline. [Zaitsev, 2000, p. 222]

Thus, in the West also, mathematics became once again mixed with philosophy, but this time with the philosophy of Christianity.⁹ Zaitsev also notes (2000, p. 223) that the idea of multiple layers of meaning was dear to the authors of medieval texts; but in contrast to biblical commentaries, which were strictly separated from the texts used as a source, commentaries and sources were routinely intermixed in the geometric work.

Mathematics sank to a rather low level in Europe after 500 CE, recovering only slightly if at all in the Carolingian Renaissance of the ninth century. One of the better-informed scholars of the tenth century was Gerbert of Aurillac (ca. 940–1002), who reigned as Pope Sylvester II during the last three years of his life. Even though Gerbert was one of the leading scholars of his day, who advocated the use of Hindu–Arabic numerals, one of his letters to a certain Adalbold of Liège is occupied with a discussion of the rule for finding the area of a triangle given its base and altitude! The general level of geometry, however, was not so bad as the correspondence between Gerbert and Adalbold seems to imply. In fact, Gerbert wrote, but did not finish, *Geometria*, a practical manual of surveying based on what was probably in Boethius' textbook. This work, which can be read online,¹⁰ consists of 89 brief chapters devoted to triangles, circles, spheres, and regular polygons. It gives the names of standard units of length and finds the areas of such simple figures as a trapezoid (Chapter XLVIII) and a semicircle (Chapter LXXIX, where the rule is given to multiply the square of the diameter by 11 and divide the product by 28). A specimen that may be typical of the level of geometric knowledge used in civil engineering, architecture, surveying, and geometric astronomy in the twelfth century, just as translations from Arabic works began to circulate in Europe, is provided by Hugh of St. Victor's *Practical Geometry* (Homann, 1991), in which one can find a description of the construction and use of an astrolabe (a fundamental tool used by

⁹ Compare with the quotations from the pseudo-Boethius and Gerbert in Chapter 3.

¹⁰ <http://p1d.chadwyck.com>, a commercial site that provides the *Patrologia Latina* of Jacques-Paul Migne (1800–1875). Search for the title “geometria.”

navigators and explorers for many centuries)¹¹ and a discussion of different ways of using similar triangles to determine distances to inaccessible objects.

3.1. Late Medieval and Renaissance geometry. In the late Middle Ages Europeans from many countries eagerly sought Latin translations of Arabic treatises, just as some centuries earlier Muslim scholars had sought translations of Hindu and Greek treatises. In both cases, those treatises were made the foundation for ever more elaborate and beautiful mathematical theories. An interesting story arises in connection with these translations, showing the unreliability of transmission. The Sanskrit word “bowstring” (*jya*) used for a half-chord of a circle was simply borrowed by the Arab translators and written as *jb*, apparently pronounced *jiba*, since Arabic was written without vowels. Over time, this word came to be interpreted as *jaib*, meaning a pocket or fold in a garment. When the Arabic works on trigonometry were translated into Latin in the twelfth century, this word was translated as *sinus*, which also means a pocket or cavity. The word caught on very quickly, apparently because of the influence of Leonardo of Pisa, and is now well established in all European languages. That is the reason we now have three trigonometric functions, the secant (Latin for cutting), the tangent (Latin for touching), and the sine (Latin for a concept having nothing at all to do with geometry!).

We think of analytic geometry as the application of algebra to geometry. Its origins in Europe, however, antedate the high period of European algebra by a century or more. The first adjustment in the way mathematicians think about physical dimensions, an essential step on the way to analytic geometry, occurred in the fourteenth century.

Nicole d'Oresme. The first prefiguration of analytic geometry occurs in the work of Nicole d'Oresme (1323–1382). The *Tractatus de latitudinibus formarum*, published in Paris in 1482 and ascribed to Oresme but probably written by one of his students, contains descriptions of the graphical representation of intensities. The crucial realization that he came to was that since the area of a rectangle is computed by multiplying length and width and the distance traveled at constant speed is computed by multiplying velocity and time, it follows that if one line is taken proportional to time and a line perpendicular to it is proportional to a (constant) velocity, the area of the resulting rectangle is proportional to the distance traveled.

Oresme considered three forms of qualities, which he labeled *uniform*, *uniformly difform*, and *difformly difform*. We would call these classifications constant, linear, and nonlinear. Examples are provided in Fig. 6, although Oresme realized that the “difformly difform” constituted a large class of qualities and mentioned specifically that a semicircle could be the representation of such a quality.

The advantage of representing a *distance* by an *area* rather than a line appeared in the case when the velocity changed during a motion. In the simplest nontrivial case the velocity was uniformly difform. In that case, the distance traversed is what it would have been had the body moved the whole time with the velocity it

¹¹ The French explorer Samuel de Champlain (1567–1635) apparently lost his astrolabe while exploring the Ottawa River in 1613. Miraculously, that astrolabe was found 254 years later, in 1867, and the errors in Champlain's diaries were used by an author named Alex Jamieson Russell (1807–1887) to establish the fact and date of the loss (Russell, 1879). The discovery of this astrolabe only a month after the founding of the Canadian Federation was of metaphorical significance to Canadian poets. See, for example, *The Buried Astrolabe*, by Craig Stewart Walker, McGill–Queens University Press, Montreal, 2001, a collection of essays on Canadian dramatists.

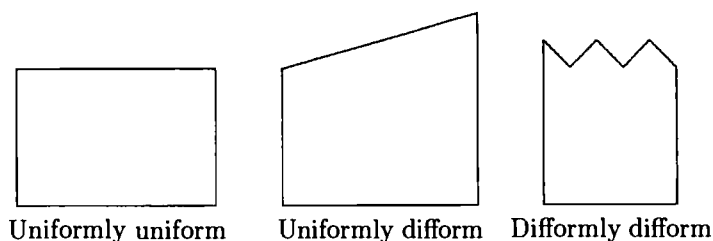


FIGURE 6. Nicole Oresme's classification of motions.

had at the midpoint of the time of travel. This is the case now called *uniformly accelerated* motion. According to Clagett (1968, p. 617), this rule was first stated by William Heytesbury (ca. 1313–ca. 1372) of Merton College, Oxford around 1335 and was well known during the Middle Ages.¹² It is called the *Merton Rule*. In his book *De configurationibus qualitatum et motuum*, Oresme applied these principles to the analysis of such motion and gave a simple geometric proof of the Merton Rule. He illustrated the three kinds of motion by drawing a figure similar to Fig. 6. He went on to say that if a difformly difform quality was composed of uniform or uniformly difform parts, as in the example in Fig. 6, its quantity could be measured by (adding) its parts. He then pushed this principle to the limit, saying that if the quality was difform but not made up of uniformly difform parts, say being represented by a curve, then “it is necessary to have recourse to the mutual measurement of curved figures” (Clagett, 1968, p. 410). This statement must mean that the distance traveled is the “area under the velocity curve” in all three cases. Oresme unfortunately did not give any examples of the more general case, but he could hardly have done so, since the measurement of figures bounded by curves was still very primitive in his day.

Trigonometry. Analytic geometry would be unthinkable without plane trigonometry. Latin translations of Arabic texts of trigonometry, such as those of al-Tusi and al-Jayyani, which will be discussed below, began to circulate in Europe in the late Middle Ages. These works provided the foundation for such books as *De triangulis omnimodis* by Regiomontanus, published in 1533, after his death, which contained trigonometry almost in the form still taught. Book 2, for example, contains as its first theorem the law of sines for plane triangles, which asserts that the sides of triangles are proportional to the sines of the angles opposite them. The main difference between this trigonometry and ours is that a sine remains a *length* rather than a *ratio*. It is referred to an *arc* rather than to an *angle*. It was once believed that Regiomontanus discovered the law of sines for spherical triangles (Proposition 16 of Book 4) as well; but we now know that this theorem was known at least 500 years earlier to Muslim mathematicians whose work Regiomontanus must have read. A more advanced book on the subject, which reworked the reasoning of Heron on the area of a triangle given its sides, was *Trigonometriæ sive de dimensione triangulorum libri quinque* (*Five Books of Trigonometry, or, On the Size of Triangles*), first published in 1595, written by the Calvinist theologian Bartholomeus Pitiscus

¹² Boyer (1949, p. 83) says that the rule was stated around this time by another fourteenth-century Oxford scholar named Richard Suiseth, known as Calculator for his book *Liber calculatorum*. Suiseth shares with Oresme the credit for having proved that the harmonic series diverges.

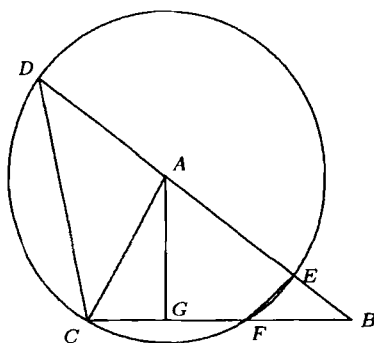


FIGURE 7. Pitiscus' derivation of the proportions in which an altitude divides a side of a triangle.

(1561–1613). This was, incidentally, the book that established the name *trigonometry* for this subject. Pitiscus showed how to determine the parts into which a side of a triangle is divided by the altitude, given the lengths of the three sides. To guarantee that the angles adjacent to the side were acute, he stated the theorem only for the altitude from the vertex of the largest angle.

Pitiscus' way of deriving this proportion was as follows. If the shortest side of the triangle ABC is AC and the longest is BC , let the altitude to BC be AG , as in Fig. 7. Draw the circle through C with center at A , so that B lies outside the circle, and let the intersections of the circle with AB and BC be E and F respectively. Then extend BA to meet the circle at D , and connect CD . Then $\angle BFE$ is the supplement of $\angle CFE$, which subtends the arc EDC , which in turn is an arc of 180° plus the arc \widehat{CD} . Hence $\angle BFE$ is the complement of the angle subtended by the arc \widehat{CD} . That in turn is the angle subtended by the supplementary arc \widehat{CE} ; thus $\angle BFE = \angle CDB$, and so the triangles BCD and BEF are similar. It follows that

$$AB^2 = AC^2 + BC^2 - 2AC \cdot BC \cos(\angle ACB),$$

which is what we now know as the *law of cosines*.

Pitiscus also gave an algebraic solution of the trisection problem discovered by an earlier mathematician, Jobst Bürgi (1552–1632). The solution had been based on the fact that the chord of triple an angle is three times the chord of the angle minus the cube of the chord of the angle. This relation makes no sense in terms of geometric dimension; it is a purely numerical relation. It is interesting that it is stated in terms of chords, since Pitiscus surely knew about sines.

4. Geometry in the Muslim world

In the Western world most of the advancement of geometry in the millennium from the fall of the Western Roman Empire to the fall of the Eastern Empire occurred among the Muslim and Jewish mathematicians of Baghdad, Samarkand, Cordoba, and other places. This work had some features of Euclid's style and some of Heron's. Matvievskaia (1999) has studied the extensive commentaries on the tenth book of Euclid's *Elements* written by Muslim scholars from the ninth through twelfth centuries and concluded that while formally preserving a Euclidean

distinction between magnitude and number, they actually operated with quadratic and quartic irrationals as if they were numbers.

4.1. The parallel postulate. The Islamic mathematicians continued the later Hellenistic speculation on Euclid's parallel postulate. According to Sabra (1969), this topic came into Islamic mathematics through a commentary by Simplicius on Book 1 of the *Elements*, whose Greek original is lost, although an Arabic translation exists. In fact, Sabra found a manuscript that contains Simplicius' attempted proof. The reworking of this topic by Islamic mathematicians consisted of a criticism of Simplicius' argument followed by attempts to repair its defects. Gray (1989, pp. 42–54) presents a number of these arguments, beginning with the ninth-century mathematician al-Gauhari. Al-Gauhari attempted to show that two lines constructed so as to be parallel, as in Proposition 27 of Book 1 of the *Elements* must also be equidistant at all points. If he had succeeded, he would indeed have proved the parallel postulate.

4.2. Thabit ibn-Qurra. Thabit ibn-Qurra (826–901), whose revision of the Arabic translation of Euclid became a standard in the Muslim world, also joined the debate over the parallel postulate. According to Gray (1989, pp. 43–44), he considered a solid body moving without rotating so that one of its points P traverses a straight line. He claimed that the other points in the body would also move along straight lines, and obviously, they would remain equidistant from the line generated by the point P . By regarding these lines as completed loci, he avoided a certain objection that could be made to a later argument of ibn al-Haytham, discussed below. Thabit ibn-Qurra's work on this problem was ground-breaking in a number of ways, anticipating much that is usually credited to the eighteenth-century mathematicians Lambert and Saccheri. He proved, for example, that if a quadrilateral has two equal adjacent angles, and the sides not common to these two angles are equal, then the other two angles are also equal to each other. In the case when the equal angles are right angles, such a figure is called—unjustly, we may say—a *Saccheri quadrilateral*, after Giovanni Saccheri (1667–1733), who like Thabit ibn-Qurra, developed it in an attempt to prove the parallel postulate. Gray prefers to call it a *Thabit quadrilateral*, and we shall use this name. Thabit ibn-Qurra's proof amounted to the claim that a perpendicular drawn from one leg of such a quadrilateral to the opposite leg would also be perpendicular to the leg from which it was drawn. Such a figure, a quadrilateral having three right angles, or half of a Thabit quadrilateral, is now called—again, unjustly—a *Lambert quadrilateral*, after Johann Heinrich Lambert (1728–1777), who used it for the same purpose. We should probably call it a *semi-Thabit quadrilateral*. Thabit's claim is that either type of Thabit quadrilateral is in fact a rectangle. If this conclusion is granted, it follows by consideration of the diagonals of a rectangle that the sum of the acute angles in a right triangle is a right angle, and this fact makes Thabit's proof of the parallel postulate work.

The argument of Thabit ibn-Qurra, according to Gray, is illustrated in Fig. 8.¹³ Given three lines l , m , and n such that l is perpendicular to n at E and m intersects it at A , making an acute angle, let W be any point on m above n and draw a perpendicular WZ from W to n . If E is between A and Z , then l must intersect m by virtue of what is now called *Pasch's theorem*. That much of the argument would

¹³ We are supplementing the figure and adding steps to the argument for the sake of clarity.

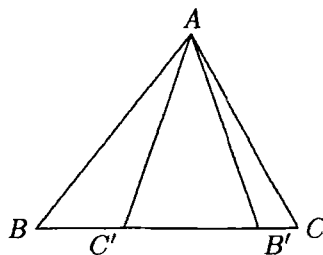


FIGURE 9. Thabit ibn-Qurra's Pythagorean theorem.

4.3. Al-Kuhi. A mathematician who devoted himself almost entirely to geometry was Abu Sahl al-Kuhi (ca. 940–ca. 1000), the author of many works, of which some 30 survive today. Berggren (1989), who has edited these manuscripts, notes that 14 of them deal with problems inspired by the reading of Euclid, Archimedes, and Apollonius, while 11 others are devoted to problems involving the compass, spherical trigonometry, and the theory of the astrolabe. Berggren presents as an example of al-Kuhi's work the angle trisection shown in Fig. 10. In that figure the angle φ to be trisected is ABG , with the base BG horizontal. The idea of the trisection is to extend side AB any convenient distance to D . Then, at the midpoint of BD , draw a new set of mutually perpendicular lines making an angle with the horizontal equal to $\varphi/2$, and draw the rectangular hyperbola through B having those lines as asymptotes. Apollonius had shown (*Conics*, Book 1, Propositions 29 and 30) that D lies on the other branch of the hyperbola. Then BE is drawn equal to BD , that is a circle through D with center at B is drawn, and its intersection with the hyperbola is labeled E . Finally, EZ is drawn parallel to BG . It then follows that $\varphi = \angle AZE = \angle ZBE + \angle ZEB = 3\theta$, as required.

4.4. Al-Haytham. One of the most prolific and profound of the Muslim mathematician-scientists was Abu Ali ibn al-Haytham (965–1040), known in the West as Alhazen. He was the author of more than 90 books, 55 of which survive.¹⁴ A significant indication of his mathematical prowess is that he attempted to reconstruct the lost Book 8 of Apollonius' *Conics*. His most famous book is his *Treatise on Optics* (*Kitab al-Manazir*) in seven volumes. The fifth volume contains the problem known as Alhazen's problem: *Given the location of a surface, an object, and an observer, find the point on the surface at which a light ray from the object will be reflected to the observer.* Rashed (1990) points out that burning-mirror problems of this sort had been studied extensively by Muslim scholars, especially by Abu Saad ibn Sahl some decades before al-Haytham. More recently (see Guizal and Dudley, 2002) Rashed has discovered a manuscript in Teheran written by ibn Sahl containing precisely the law of refraction known in Europe as *Snell's law*, after Willebrod Snell (1591–1626) or *Descartes' law*.¹⁵ The law of refraction as given by Ptolemy in the form of a table of values of the angle of refraction and the angle of incidence implied that the angle of refraction was a quadratic function of the angle of incidence. The actual relation is that the ratio of the sines of the two angles is a constant for refraction at the interface between two different media.

¹⁴ Rashed (1989) suggested that these works and the biographical information about al-Haytham may actually refer to two different people. The opposite view was maintained by Sabra (1998).

¹⁵ According to Guizal and Dudley, this law was stated by Thomas Harriot (1560–1621) in 1602.

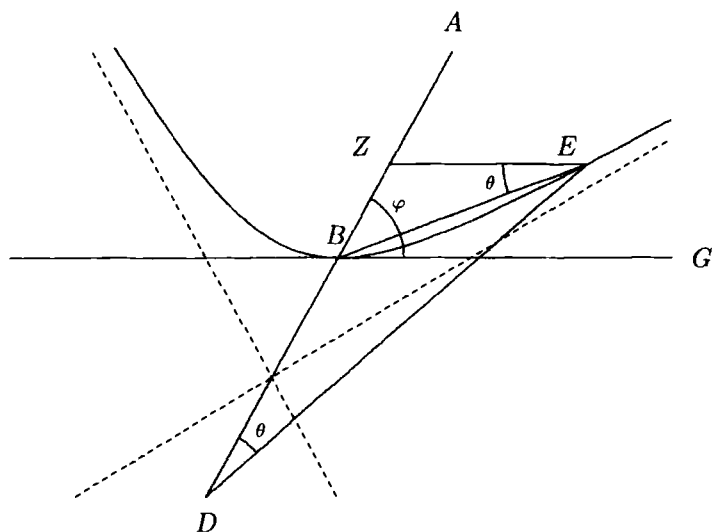


FIGURE 10. Al-Kuhi's angle trisection.

Al-Haytham also attempted to prove the parallel postulate. According to Gray (1989, p. 45), the argument given by al-Haytham in his *Commentary on the Premises to Euclid's Book The Elements*, and later in his *Book on the Resolution of Doubts* was based on the idea of translating a line perpendicular to a given line in such a way that it always remains perpendicular. The idea is that the endpoint of the line must trace a straight line parallel to the directing line.

4.5. Omar Khayyam. In his paper "Discussion of difficulties in Euclid" (Amir-Moez, 1959), the Persian mathematician Omar Khayyam (1048–1131) raised a number of questions about al-Haytham's argument. He asked how a line could move while remaining perpendicular to a given line, and more generally, how geometry and motion could be connected. Even admitting that Euclid allowed a line to be generated by a moving point and a surface by a moving line, he pointed out that al-Haytham was requiring something more in demanding that one line remain perpendicular to another at each instant during its motion.¹⁶

Having refuted al-Haytham's proof, Omar Khayyam himself attempted a proof (Amir-Moez, 1959) based on a proposition that he claimed Aristotle had proved: *If two lines converge, they will (eventually) intersect*. This claim raises an interesting question, since as we have seen, Aristotle did not accept the arguments given by scholars in Plato's Academy to prove that parallel lines exist. Given his disbelief in a completed infinity, he probably would have liked an argument proving that

¹⁶ Omar Khayyam's objection is right on target from the point of view of modern physics. If the special theory of relativity is correct, no sense can be attached to the statement that two events occurring at different places are simultaneous. One observer may find them so, while another does not agree. The same objection applies to Thabit ibn-Qurra's argument, which assumes a rigid body. In special relativity rigid bodies do not exist. What al-Haytham did was to ignore all points from the moving solid except those lying along a certain line. The relation between motion and geometry lies at the heart of relativity theory.

converging lines must intersect. Although none of the writings now attributed to Aristotle contain such an argument, Gray (1989, p. 47) points out that Omar Khayyam may have had access to Aristotelian treatises that no longer exist. In any case, he concluded on the basis of Aristotle's argument that two lines that converge on one side of a transversal must diverge on the other side. With that, having proved correctly that the perpendicular bisector of the base of a Thabit quadrilateral is also the perpendicular bisector of the summit, Omar Khayyam concluded that the base and summit could not diverge on either side, and hence must be equidistant. Like Thabit ibn-Qurra's proof, his proof depended on building one Thabit quadrilateral on top of another by doubling the common bisector of the base and summit, then crossing its endpoint with a perpendicular which (he said) would intersect the extensions of the lateral sides. Unfortunately, if that procedure is repeated often enough in hyperbolic geometry, those intersections will not occur.

All of these mathematicians were well versed in the Euclidean tradition of geometry. In the preface to his book on algebra, Omar Khayyam says that no one should attempt to read it who has not already read Euclid's *Elements* and *Data* and the first two books of Apollonius' *Conics*. His reason for requiring this background was that he intended to use conic sections to solve cubic and quartic equations geometrically. This book contains Euclidean rigor attached to algebra in a way that fits equally well into the history of both algebra and geometry. However, in other places it seems clear that Omar Khayyam was posing geometric problems for the sake of getting interesting equations to solve. For example (Amir-Moez, 1963), he posed the problem of finding a point on a circle such that the perpendicular from the point to a radius has the same ratio to the radius that the two segments into which it divides the radius have to each other. If the radius is r and the length of the longer segment cut off on the radius is the unknown x , the equation to be satisfied is $x^3 + rx^2 + r^2x = r^3$. Without actually writing out this equation, Omar Khayyam showed that the geometric problem amounted to using the stated condition to find the second asymptote of a rectangular hyperbola, knowing one of its asymptotes and one point on the hyperbola. However, he regarded that analysis as merely an introduction to his real purpose, which was a discussion of the kinds of cubic equations that require conic sections for their solution. After a digression to classify these equations, he returned to the original problem, and finally, showed how to solve it using a rectangular hyperbola. He found the arc to be about 57° , so that $x \approx r \cos(57^\circ) = 0.544r$. Omar Khayyam described x as being about $30\frac{2}{3}$ pieces, that is, sixtieths of the radius. We reserve the discussion of this combination of algebra and geometry for Chapter 14.

As his work on the parallel postulate shows, Omar Khayyam was very interested in logical niceties. In the preface to his *Algebra* and elsewhere (for example, Amir-Moez, 1963, p. 328) he shows his adherence to Euclidean standards, denying the reality of a fourth dimension:

If the algebraist were to use the square of the square in measuring areas, his result would be figurative [theoretical] and not real, because it is impossible to consider the square of the square as a magnitude of a measurable nature. . . This is even more true in the case of higher powers. [Kasir, 1931, p. 48]

4.6. Nasir al-Din al-Tusi. The thirteenth century was as disruptive to the Islamic world as the fifth century had been to the Roman world. This was the time

of the Mongol expansion, which brought the conquest of China in the early part of the century, then the conquest of Kievan Rus in 1243, and finally, the sack of Baghdad in 1254. Despite the horrendous times, the astronomer-mathematician Nasir al-Din al-Tusi (1201–1274) managed to produce some of the best mathematics of the era. Al-Tusi was treated with respect by the Mongol conqueror of Baghdad, who even built for him an astronomical observatory, at which he made years of accurate observations and improved the models in Ptolemy's *Almagest*.¹⁷ Al-Tusi continued the Muslim work on the problem of the parallel postulate. According to Gray (1989, pp. 50–51), al-Tusi's proof followed the route of proving that the summit angles of a Thabit quadrilateral are right angles. He showed by arguments that Euclid would have accepted that they cannot be obtuse angles, since if they were, the summit would diverge from the base as a point moves from either summit vertex toward the other. Similarly, he claimed, they could not be acute, since in that case the summit would converge toward the base as a point moves from either summit vertex toward the other. Having thus argued that a Thabit quadrilateral must be a rectangle, he could give a proof similar to that of Thabit ibn-Qurra to establish the parallel postulate.

In a treatise on quadrilaterals written in 1260, al-Tusi also reworked the trigonometry inherited from the Greeks and Hindus and developed by his predecessors in the Muslim world, including all six triangle ratios that we know today as the trigonometric functions. In particular, he gave the law of sines for spherical triangles, which states that the sines of great-circle arcs forming a spherical triangle are proportional to the sines of their opposite angles. According to Hairtdinova (1986) trigonometry had been developing in the Muslim world for some centuries before this time, and in fact the mathematician Abu Abdullah al-Jayyani (989–1079), who lived in the Caliphate of Cordoba, wrote *The Book on Unknown Arcs of a Sphere*, a treatise on plane and spherical trigonometry. Significantly, he treated ratios of lines as numbers, in accordance with the evolution of thought on this subject in the Muslim world. Like other Muslim mathematicians, though, he does not use negative numbers. As Hairtdinova mentions, there is clear evidence of the Muslim influence in the first trigonometry treatise written by Europeans, the book *De triangulis omnimodis* by Regiomontanus, whose exposition of plane trigonometry closely follows that of al-Jayyani.

Among these and many other discoveries, al-Tusi discovered the interesting theorem that if a circle rolls without slipping inside a circle twice as large, each point on the smaller circle moves back and forth along a diameter of the larger circle. This fact is easy to prove and an interesting exercise in geometry. It has obvious applications in geometric astronomy, and was rediscovered three centuries later by Copernicus and used in Book 3, Chapter 4 of his *De revolutionibus*.

5. Non-Euclidean geometry

The centuries of effort by Hellenistic and Islamic mathematicians to establish the parallel postulate as a fact of nature began to be repeated in early modern Europe with the efforts of a number of mathematicians to replace the postulate with some other assumption that seemed indubitable. Then, around the year 1800, a change

¹⁷ The world's debt to Muslim astronomers is shown in the large number of stars bearing Arabic names, such as Aldebaran (the Follower), Altair (the Flyer), Algol (the Ghoul), Betelgeuse (either the Giant's Hand or the Giant's Armpit), and Deneb (the Tail).

in attitude took place, as a few mathematicians began to explore non-Euclidean geometries as if they might have some meaning after all. Within a few decades the full light of day dawned on this topic, and by the late nineteenth century, models of the non-Euclidean geometries inside Euclidean and projective geometry removed all doubt as to their consistency. This history exhibits a sort of parallelism with the history of the classical construction problems and with the problem of solving higher-degree equations in radicals, all of which were shown in the early nineteenth century to be impossible tasks. In all cases, the result was a deeper insight into the original questions. In all three cases, group theory came to play a role, although a much smaller one in the case of non-Euclidean geometry than in the other two.

5.1. Girolamo Saccheri. The Jesuit priest Girolamo Saccheri (1667–1733), a professor of mathematics at the University of Pavia, published in the last year of his life the treatise *Euclides ab omni nævo vindicatus* (*Euclid Acquitted of Every Blemish*), a good example of the creativity a very intelligent person will exhibit when trying to retain a strongly held belief. Some of his treatise duplicates what had already been done by the Islamic mathematicians, including the study of Thabit quadrilaterals, that is, quadrilaterals having a pair of equal opposite sides and equal base angles and also quadrilaterals having three right angles. Saccheri deduced with strict rigor all the basic properties of Thabit quadrilaterals with right angles at the base.¹⁸ He realized that the fundamental question involved the summit angles of these quadrilaterals—Saccheri quadrilaterals, as they are now called. Since these angles were equal, the only question was whether they were obtuse, right, or acute angles. He showed in Propositions 5 and 6 that if one such quadrilateral had obtuse summit angles, then all of them did likewise, and that if one had right angles, then all of them did likewise. It followed by elimination and without further proof (Proposition 7, which Saccheri proved anyway) that if one of them had acute angles, then all of them did likewise. Not being concerned to eliminate the possibility of the right angle, which he believed was the true one, he worked to eliminate the other two hypotheses.

He showed that the postulate as Euclid stated it is true under the hypothesis of the obtuse angle. That is, two lines cut by a transversal in such a way that the interior angles on one side are less than two right angles will meet on that side of the transversal. As we now know, that is because they will meet on *both* sides of the transversal, assuming it makes sense to talk of opposite sides. Saccheri remarked that the intersection must occur at a finite distance. This remark seems redundant, since all distances in geometry were finite until projective geometers introduced points at infinity. But Saccheri, in the end, would be reasoning about points at infinity as if something were known about them, even though he had no careful definition of them.

It is true, as many have pointed out, that his proof of this fact uses the exterior angle theorem (Proposition 16 of Book 1 of Euclid) and hence assumes that lines are infinite.¹⁹ But Euclid himself, at least as later edited, states explicitly that

¹⁸ It is unlikely that Saccheri knew of the earlier work by Thabit ibn-Qurra and others. Although Arabic manuscripts stimulated a revival of mathematics in Europe, they were apparently soon forgotten as Europeans began writing their own treatises. Coolidge (1940) gives the history of the parallel postulate jumping directly from Proclus and Ptolemy to Saccheri, never mentioning any of the Muslim mathematicians.

¹⁹ Actually, the use of that proposition is confined to elaborations by the modern reader. The proof stated by Saccheri uses only the fact that lines are *unbounded*, that is, can be extended to

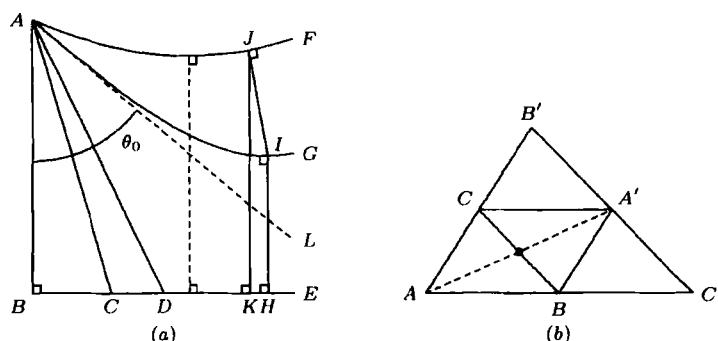


FIGURE 11. (a) Lines through A that intersect BE and those that share a common perpendicular with BE are separated by a line (AL) that is asymptotic to BE . (b) The angle defect of $\triangle AB'C'$ is more than twice the defect of $\triangle ABC$.

two lines cannot enclose an area, so that Saccheri can hardly be faulted for dealing with only one Euclidean postulate at a time. Since the parallel postulate implies that the summit and base of a Saccheri quadrilateral must meet on *both* sides of the quadrilateral under the hypothesis of the obtuse angle, even a severe critic should be inclined to give Saccheri a passing grade when he rejects this hypothesis.

Having disposed of the hypothesis of the obtuse angle, Saccheri then joined battle (his phrase) with the hypothesis of the acute angle. Here again, he proved some basic facts about what we now call hyperbolic geometry. Given any quadrilateral having right angles at the base and acute angles at the summit, it follows from continuity considerations that the length of a perpendicular dropped from the summit to the base must reach a minimum at some point, and at that point it must also be perpendicular to the summit. Saccheri analyzed this situation in detail, describing in the process a great deal of what must occur in what is now called hyperbolic geometry. In terms of Fig. 11(a),²⁰ he considered all the lines like AF through the point A such that angle BAF is acute. He wished to show that they all intersected the line BE .

Saccheri proved that there must be at least one angle θ_0 for which the line AL making that angle neither intersects BE nor has a common perpendicular with it. This line, as Saccheri showed in Proposition 23, must approach BE asymptotically as we would say. At that point he made the small slip that had been warned against even in ancient times, assuming that "approaching" implies "meeting." His intuition for hyperbolic geometry was very good, as he imagined a line perpendicular to BE moving away from AB and the lines from A perpendicular to it rotating

any length. It is not necessary to require that the extension never overlap the portion already present.

²⁰ Since the flat page is not measurably non-Euclidean, and wouldn't be even if spread out to cover the entire solar system, the kinds of lines that occur in hyperbolic geometry cannot be drawn accurately on paper. Our convention is the usual one: When asymptotic properties are not involved, draw the lines straight. When asymptotic properties need to be shown, draw them as hyperbolas. Actually, if the radius of curvature of the plane were comparable to the width of the page, two lines with a common perpendicular would diverge from each other like the graphs of $\cosh x$ and $-\cosh x$, very rapidly indeed.

clockwise about A to make angles that decreased to θ_0 . He then—too hastily, as we now know—drew the conclusion that θ_0 would have the properties of *both* of the sets of angles that it separated, that is, the line making this angle would intersect BE and would also have a common perpendicular with it. In fact, it has neither property. But Saccheri was determined to have both. As he described the situation, the hypothesis of the acute angle implied the existence of two straight lines that have a common perpendicular *at the same point*. In other words, there could be two distinct lines perpendicular to the same line at a point, which is indeed a contradiction. Unfortunately, the point involved was not a point of the plane, but is infinitely distant, as Saccheri himself realized. But he apparently believed that points and lines at infinity must obey the same axioms as those in the finite plane.

Once again, as in the case of Ptolemy, Thabit ibn-Qurra, ibn al-Haytham, and others, Saccheri had developed a new kind of geometry, but resorted to procrustean methods to reconcile it with the geometry he believed in.

5.2. Lambert and Legendre. The writings of the Swiss mathematician Johann Heinrich Lambert (1728–1777) seem modern in many ways. For example, he proved that π is irrational (specifically, that $\tan x$ and x cannot both be rational numbers), studied the problem of constructions with straightedge and a fixed compass, and introduced the hyperbolic functions and their identities as they are known today, including the notation $\sinh x$ and $\cosh x$. He wrote, but did not publish, a treatise on parallel lines, in which he pointed out that the hypothesis of the obtuse angle holds for great circles on a sphere and that the area of a spherical triangle is the excess of its angle sum over π times the square of the radius. He concluded that in a sphere of imaginary radius ir , whose area would be negative, the area of a triangle might be proportional to the excess of π over the angle sum. What a sphere of imaginary radius looks like took some time to discern, a full century, to be exact.

By coincidence, the hyperbolic functions that he studied turned out to be the key to trigonometry in this imaginary world. Just as on the sphere there is a natural unit of length (the radius of the sphere, for example), the same would be true, as Lambert realized, on his imaginary sphere. Such a unit could be selected in a number of ways. The angle θ_0 mentioned above, for example, decreases steadily as the length AB increases. Hence every length is associated with an acute angle, and a natural unit of length might be the one associated with half of a right angle. Or, it might be the length of the side of an equilateral triangle having a specified angle. In any case, Lambert at least recognized that he had not proved the parallel postulate. As he said, it was always possible to develop a proof of the postulate to the point that only some small, seemingly obvious point remained unproved, but that last point nearly always concealed an assumption equivalent to what was being proved.

Some of Lambert's reasoning was recast in more precise form by Legendre, who wrote a textbook of geometry used in many places during the nineteenth century, including (in English translation) the United States. Legendre, like Lambert and Saccheri, refuted the possibility that the angle sum of a triangle could be more than two right angles and attempted to show that it could not be less. Since the defect of a triangle—the difference between two right angles and its angle sum—is additive, in the sense that if a triangle is cut into two smaller triangles, the defect of the larger triangle is the sum of the defects of the two smaller ones, he saw correctly that if one could repeatedly double a triangle, eventually the angle sum would have

to become negative, which was surely impossible. Unfortunately, the possibility of repeated doubling that he had in mind was just one of those small points mentioned by Lambert that turn out to be equivalent to the parallel postulate. In fact, it is rather easy to see that such is the case, since (Fig. 11(b)) the possibility of drawing a line $B'C'$ through a point A' inside the angle CAB that intersects both AB and AC is simply another way of saying that the lines AB and AC must both intersect *some* line through A' , that is, AC cannot be parallel to *every* line through A' that intersects AB .

5.3. Gauss. The true situation in regard to the parallel postulate was beginning to be understood by the end of the eighteenth century. Gauss, who read Lambert's work on parallels (which had been published posthumously), began to explore this subject as a teenager, although he kept his thoughts to himself except for letters to colleagues and never published anything on the subject. His work in this area was published in Vol. 8 of the later edition of his collected works. It is nicely summarized by Klein (1926, pp. 58–59). In 1799 he wrote to Farkas Bolyai (1775–1856), his classmate from Göttingen, that he could prove the parallel postulate provided that triangles of arbitrarily large area were admitted. Such a confident statement can only mean that he had developed the metric theory of hyperbolic geometry to a considerable extent. Five years later he wrote again to explain the error in a proof of the parallel postulate proposed by Bolyai. Gauss, like Lambert, realized that a non-Euclidean space would have a natural unit of length, and mentioned this fact in a letter of 1816 to his student Christian Ludwig Gerling (1788–1864), proposing as unit the side of an equilateral triangle whose angles were $59^\circ 59' 59.99999 \dots''$.²¹ To Gauss' surprise, in 1818 he received from Gerling a paper written by Ferdinand Karl Schweikart (1780–1859), a lawyer then in Marburg, who had developed what he called *astral geometry*. It was actually hyperbolic geometry, and Schweikart had gone far into it, since he knew that there was an upper bound to the area of a triangle in this geometry, that its metric properties depended on an undetermined constant C (its radius of curvature), and that it contained a natural unit of length, which he described picturesquely by saying that if that length were the radius of the earth, then the line joining two stars would be tangent to the earth. Gauss wrote back to correct some minor points of bad drafting on Schweikart's part (for example, Schweikart neglected to say that the stars were assumed infinitely distant), but generally praising the work. In fact, he communicated his formula for the limiting area of a triangle:

$$\frac{\pi C^2}{(\ln(1 + \sqrt{2}))^2}.$$

By coincidence, Schweikart's nephew Franz Adolph Taurinus (1794–1874), also a lawyer, who surely must have known of his uncle's work in non-Euclidean geometry, sent Gauss his attempt at a proof of the parallel postulate in 1824. Gauss explained the true situation to Taurinus under strict orders to keep the matter secret. The following year, Taurinus published a treatise *Geometriæ prima elementa* (*First Elements of Geometry*) in which he accepted the possibility of other geometries. Gauss wrote to the astronomer-mathematician Friedrich Wilhelm Bessel

²¹ In comparison with the radius of curvature of space, this would be an extremely small unit of length; however, if space is curved negatively at all, its radius of curvature is so enormous that in fact this unit might be very large.

(1784–1846) in 1829 that he had been thinking about the foundations of geometry off and on for nearly 40 years (in other words, from the age of 13 on), saying that his investigations were “very extensive,” but probably wouldn’t be published, since he feared the controversy that would result. Some time during the mid-1820s, the time when he was writing and publishing his fundamental work on differential geometry, Gauss wrote a note—which, typically, he never published—in which he mentioned that revolving a tractrix about its asymptote produced a surface that is the opposite of a sphere. This surface turns out to be a perfect local model of the non-Euclidean geometry in which the angle sum of a triangle is less than two right angles. It is now called a pseudosphere. This same surface was discussed a decade later by Ferdinand Minding (1806–1885), who pointed out that some pairs of points on this surface can be joined by more than one minimal path, just like antipodal points on a sphere.

5.4. Lobachevskii and János Bolyai. From what has been said so far, it is clear that the full light of day was finally dawning on the subject of the parallel postulate. As more and more mathematicians worked over the problem and came to the same conclusion, from which others gained insight little by little, all that remained was a slight push to tip the balance from attempts to prove the parallel postulate to the exploration of alternative hypotheses. The fact that this extra step was taken by several people nearly simultaneously can be expressed poetically, as it was by Felix Klein (1926, p. 57), who referred to “one of the remarkable laws of human history, namely that the times themselves seem to hold the great thoughts and problems and offer them to heads gifted with genius when they are ripe.” But we need not be quite so lyrical about a phenomenon that is entirely to be expected: When many intelligent people who have received similar educations work on a problem, it is quite likely that more than one of them will make the same discovery.

The credit for first putting forward hyperbolic geometry for serious consideration must belong to Schweikart, since Gauss was too reticent to do so. However, credit for the first full development of it, including its trigonometry, is due to the Russian mathematician Nikolai Ivanovich Lobachevskii (1792–1856) and the Hungarian János Bolyai (1802–1860), son of Farkas Bolyai. Their approaches to the subject are very similar. Both developed the geometry of the hyperbolic plane and then extended it to three-dimensional space. In three-dimensional space they considered the entire set of directed lines parallel to a given directed line in a given direction. Then they showed that a surface (now called a *horosphere*) that cuts all of these lines at right angles has all the properties of a Euclidean plane. By studying sections of this surface they were able to deduce the trigonometry of their new geometry. In modern terms the triangle formulas fully justify Lambert’s assertion that this kind of geometry is that of a sphere of imaginary radius. Here, for example, is the Pythagorean theorem for a right triangle of sides a , b , c in spherical and hyperbolic geometry, derived by both Lobachevskii and Bolyai, but not in the notation of hyperbolic functions. Since $\cos(ix) = \cosh(x)$ the hyperbolic formula can be obtained from the spherical formula by replacing the radius r with ir , just as Lambert stated.

Spherical geometry	Hyperbolic geometry
$\cos\left(\frac{a}{r}\right)\cos\left(\frac{b}{r}\right) = \cos\left(\frac{c}{r}\right)$	$\cosh\left(\frac{a}{r}\right)\cosh\left(\frac{b}{r}\right) = \cosh\left(\frac{c}{r}\right).$

Lobachevskii's geometry. Lobachevskii connected the parts of a hyperbolic triangle through his formula for the angle of parallelism, which is the angle θ_0 referred to above, as a function of the length AB . He gave this formula as

$$\tan\left(\frac{1}{2}F(\alpha)\right) = e^\alpha,$$

where α denotes the length AB and $F(\alpha)$ the angle θ_0 . Here e could be any positive constant, since the radius of curvature of the hyperbolic plane could not be determined. However, Lobachevskii found it convenient to take this constant to be $e = 2.71828\dots$. In effect, he took the radius of curvature of the plane as the unit of length. Lobachevskii gave the Pythagorean theorem, for example, as

$$\sin F(a) \sin F(b) = \sin F(c).$$

Of the two nearly simultaneous creators of hyperbolic geometry and trigonometry, Lobachevskii was the first to publish, unfortunately in a journal of limited circulation. He was a professor at the provincial University of Kazan' in Russia and published his work in 1826 in the proceedings of the Kazan' Physico-Mathematical Society. He reiterated this idea over the next ten years or so, developing its implications. Like Gauss, he drew the conclusion that only observation could determine if actual space was Euclidean or not. As luck would have it, the astronomers were just beginning to attempt measurements on the interstellar scale. In particular, by measuring the angles formed by the lines of sight from the Earth to a given fixed star at intervals of six months, one could get the base angles of a gigantic triangle and thereby (since the angle sum could not be larger than two right angles, as everyone agreed) place an upper bound on the size of the parallax of the star (the angle subtended by the Earth's orbit from that star). Many encyclopedias claim that the first measurement of stellar parallax was carried out in Königsberg by Bessel in 1838, and that he determined the parallax of 61 Cygni to be 0.3 seconds. Russian historians credit another Friedrich Wilhelm, namely Friedrich Wilhelm Struve (1793–1864), who emigrated to Russia and is known there as Vasilii Yakovlevich Struve. He founded the Pulkovo Observatory in 1839. Struve determined the parallax of the star Vega in 1837. Attempts to determine stellar parallax must have been made earlier, since Lobachevskii cited such measurements in an 1829 work and claimed that the measured parallax was less than $0.000372''$, which is much smaller than any observational error.²² As he said (see his collected works, Vol. 1, p. 207, quoted by S. N. Kiro, 1967, Vol. 2, p. 159):

At the very least, astronomical observations prove that all the lines amenable to our measurements, even the distances between celestial bodies, are so small in comparison with the length taken as a unit in our theory that the equations of (Euclidean) plane trigonometry, which have been used up to now must be true without any sensible error.

²² The vast distances between stars make terrestrial units of length inadequate. The light-year (about $9.5 \cdot 10^{12}$ km) is the most familiar unit now used, particularly good, since it tells us "what time it was" when the star emitted the light we are now seeing. Stellar parallax provides another unit, the parsec, which is the distance at which the radius of the Earth's orbit subtends an angle of $1''$. A parsec is about 3.258 light-years.

Thus, ironically, the acceptance of the logical consistency of hyperbolic geometry was accompanied by a nearly immediate rejection of any practical application of it in astronomy or physics. That situation was to change only much later, with the advent of relativity.

Lobachevskii was unaware of the work of Gauss, since Gauss kept it to himself and urged others to do likewise. Had Gauss been more talkative, Lobachevskii would easily have found out about his work, since his teacher Johann Martin Christian Bartels (1769–1836) had been many years earlier a teacher of the 8-year-old Gauss and had remained a friend of Gauss. As it was, however, although he continued to perfect his “imaginary geometry,” as he called it, and wrote other mathematical papers, he made his career in administration, as rector of Kazan’ University. He at least won some recognition for his achievement during his lifetime, and his writings were translated into French and German after his death and highly regarded.

Even though his imaginary geometry was not used directly to describe the world, Lobachevskii found some uses for it in providing geometric interpretations of formulas in analysis. In particular, his paper “Application of imaginary geometry to certain integrals,” which he published in 1836, was translated into German in 1904, with its misprints corrected (Liebmann, 1904). Just as we can compute the seemingly complicated integral

$$\int_0^r \sqrt{r^2 - x^2} dx = \frac{\pi}{4} r^2$$

immediately by recognizing that it represents the area of a quadrant of a circle of radius r , he could use the differential form for the element of area in rectangular coordinates in the hyperbolic plane given by $dS = (1/\sin y') dx dy$, where y' is the angle of parallelism for the distance y (in our terms $\sin y' = \operatorname{sech} y$) to express certain integrals as the non-Euclidean areas of simple figures. In polar coordinates the corresponding element of area is $dS = \cot r' dr d\theta = \sinh r dr d\theta$. Lobachevskii also gave the elements of volume in rectangular and spherical coordinates and computed 49 integrals representing hyperbolic areas and volumes, including the volumes of pyramids. These volumes turn out to involve some very complicated integrals indeed. He proved, for example, that

$$\int_0^\pi \int_0^\infty (e^x - e^{-x}) F' [a(e^x + e^{-x}) + b \cos \omega (e^x - e^{-x})] dx d\omega = \frac{-\pi}{\sqrt{a^2 - b^2}} F [2\sqrt{a^2 - b^2}].$$

Bolyai's fate. János Bolyai's career turned out less pleasantly than Lobachevskii's. Even though he had the formula for the angle of parallelism in 1823, a time when Lobachevskii was still hoping to vindicate the parallel postulate, he did not publish it until 1831, five years after Lobachevskii's first publication. Even then, he had only the limited space of an appendix to his father's textbook to explain himself. His father sent the appendix to Gauss for comments, and for once Gauss became quite loquacious, explaining that he had had the same ideas many years earlier, and that none of these discoveries were new to him. He praised the genius of the young Lobachevskii for discovering it, nevertheless. Bolyai the younger was not overjoyed at this response. He suspected Gauss of trying to steal his ideas. According to Paul Stäckel (1862–1919), who wrote the story of the Bolyais, father and son (quoted in Coolidge, 1940, p. 73), when Lobachevskii's work began to be known, Bolyai immediately thought that Gauss was stealing his work and publishing it under the

pseudonym Lobachevskii, since "it is hardly likely that two or even three people knowing nothing of one another would produce almost the same result by different routes."

5.5. The reception of non-Euclidean geometry. Some time was required for the new world revealed by Lobachevskii and Bolyai to attract the interest of the mathematical community. Because it seemed possible—even easy—to prove that parallel lines exist, or equivalently, that the sum of the angles of a triangle could not be *more* than two right angles, one can easily understand why a sense of symmetry would lead to a certain stubbornness in attempts to refute the opposite hypothesis as well. Although Gauss had shown the way to a more general understanding with the concept of curvature of a surface, which could be either negative or positive, in the 1825 paper on differential geometry (published in 1827, to be discussed in detail in Sect. 3 of Chapter 12), it took Riemann's inaugural lecture in 1854 (published in 1867, also discussed in detail in Sect. 3 of Chapter 12), which made the crucial distinction between the unbounded and the infinite, to give the proper perspective. After that, acceptance of non-Euclidean geometry was quite rapid. In 1868, the year after the publication of Riemann's lecture, Eugenio Beltrami (1835–1900) realized that Lobachevskii's theorems provide a model of the Lobachevskii–Bolyai plane in a Euclidean disk. This model is described by Gray (1989, p. 112), as follows.

Imagine a directed line perpendicular to the Lobachevskii–Bolyai plane in Lobachevskii–Bolyai three-dimensional space. The entire set of directed lines that are parallel (asymptotic) to this line on the same side of the plane generates a unique horosphere tangent to the plane at its point of intersection with the line. Some of the lines parallel to the given perpendicular in the given direction intersect the original plane, and others do not. Those that do intersect it pass through the portion of the horosphere denoted Ω in Fig. 12. Shortest paths on the horosphere are obtained as its intersections with planes passing through the point at infinity that serves as its "center." These paths are called *horocycles*. But there is only one horocycle through a given point in Ω that does not intersect a given horocycle, so that the geometry of Ω is Euclidean. As a result, we have a faithful mapping of the Lobachevskii–Bolyai plane onto the interior of a disk Ω in a Euclidean plane, under which lines in the plane correspond to chords on the disk. This model provides an excellent picture of points at infinity: they correspond to the boundary of the disk Ω . Lines in the plane are parallel if and only if the chords corresponding to them have a common endpoint. Lines that have a common perpendicular in the Lobachevskii–Bolyai plane correspond to chords whose extensions meet outside the circle. It is somewhat complicated to compute the length of a line segment in the Lobachevskii–Bolyai plane from the length of its corresponding chordal segment in Ω or vice versa, and the angle between two intersecting chords is not simply related to the angle between the lines they correspond to.²³ Nevertheless these computations can be carried out from the trigonometric rules given by Lobachevskii. The result is a perfect model of the Lobachevskii–Bolyai plane *within the Euclidean plane*, obtained by formally reinterpreting the words *line*, *plane*, and *angle*. If

²³ It can be shown that perpendicular lines correspond to chords having the property that the extension of each passes through the point of intersection of the tangents at the endpoints of the other. But it is far from obvious that this property is symmetric in the two chords, as perpendicularity is for lines.

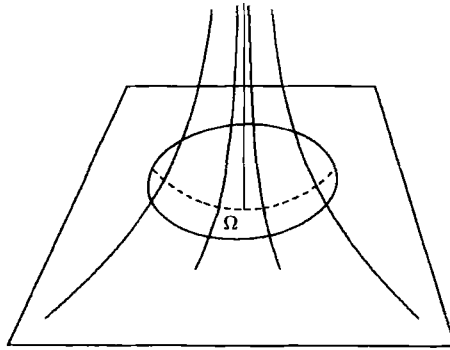


FIGURE 12. Projection of the Lobachevskii-Bolyai plane onto the interior of a Euclidean disk.



FIGURE 13. The pseudosphere. Observe that it has no definable curvature at its cusp. Elsewhere its curvature is constant and negative.

there were any contradiction in the new geometry, there would be a corresponding contradiction in Euclidean geometry itself.

A variant of this model was later provided by Henri Poincaré (1854–1912), who showed that the diameters and the circular arcs in a disk that meet the boundary in a right angle can be interpreted as lines, and in that case angles can be measured in the ordinary way.

Beltrami also provided a model of a portion of the Lobachevskii-Bolyai plane that could be embedded in three-dimensional Euclidean space: the pseudosphere obtained by revolving a tractrix about its asymptote, as shown in Fig. 13.

In 1871 Felix Klein gave a discussion of the three kinds of plane geometry in his article “Über die sogenannte nicht-Euklidische Geometrie” (“On the so-called non-Euclidean geometry”), published in the *Mathematische Annalen*. In that article he gave the classification of them that now stands, saying that the points at infinity on a line were distinct in hyperbolic geometry, imaginary in spherical geometry, and coincident in parabolic (Euclidean) geometry.

The pseudosphere is not a model of the entire Lobachevskii-Bolyai plane, since its curvature has a very prominent discontinuity. The problem of finding a surface in

three-dimensional Euclidean space that was a perfect model for the Lobachevskii-Bolyai plane, in the sense that its geodesics corresponded to straight lines and lengths and angles were measured in the ordinary way, remained open until Hilbert, in an article "Über Flächen von konstanter Gaußscher Krümmung" ("On surfaces of constant Gaussian curvature"), published in the *Transactions of the American Mathematical Society* in 1901, showed that no such surface exists.

5.6. Foundations of geometry. The problem of the parallel postulate was only one feature of a general effort on the part of mathematicians to improve on the rigor of their predecessors. This problem was particularly acute in the calculus, but the parts of calculus that raised the most doubts were those that were geometric in nature. Euclid, it began to be realized, had taken for granted not only the infinitude of the plane, but also its continuity, and had not specified in many cases what ordering of points was needed on the line for a particular theorem to be true. If one attempts to prove these theorems without drawing any figures, it becomes obvious what is being assumed. It seemed obvious, for example, that a line joining a point inside a circle to a point outside the circle must intersect the circle in a point, but that fact could not be deduced from Euclid's axioms. A complete reworking of Euclid was the result, expounded in detail in Hilbert's *Grundlagen der Geometrie* (*Foundations of Geometry*), published in 1903. This book went through many editions and has been translated into English (Bernays, 1971). In Hilbert's exposition the axioms of geometry are divided into axioms of incidence, order, congruence, parallelism, and continuity, and examples are given to show what cannot be proved when some of the axioms are omitted.

One thing is clear: No new comprehensive geometries are to be expected by pursuing the axiomatic approach of Hilbert. In a way, the geometry of Lobachevskii and Bolyai was a throwback even in its own time. The development of projective and differential geometry would have provided—indeed, *did* provide—non-Euclidean geometry by a natural expansion of the study of surfaces. It was Riemann, not Lobachevskii and Bolyai, who showed the future of geometry. The real "action" in geometry since the early nineteenth century has been in differential and projective geometry. That is not to say that no new theorems can be produced in Euclidean geometry, only that their scope is very limited. There are certainly many such theorems. Coolidge, who undertook the herculean task of writing his *History of Geometric Methods* in 1940, stated in his preface that the subject was too vast to be covered in a single treatise and that "the only way to make any progress is by a rigorous system of exclusion." In his third chapter, on "later elementary geometry," he wrote that "the temptation to run away from the difficulty by not considering elementary geometry after the Greek period at all is almost irresistible." But to attempt to build an entire theory as Apollonius did, on the synthetic methods and limited techniques in the Euclidean tool kit, would be futile. Even Lobachevskii and Bolyai at least used analytic geometry and trigonometry to produce their results. Modern geometries are much more algebraic, as we shall see in Chapter 12.

6. Questions and problems

11.1. The figure used by Zenodorus at the main step in his proof of the isoperimetric inequality had been used earlier by Euclid to show that the apparent size of objects is not inversely proportional to their distance. Prove this result by referring

that is, the intersections of the sphere with planes passing through its center. Let one “line” (great circle) be the equator of the sphere. Describe the equidistant curve generated by the endpoint of a “line segment” (arc of a great circle) of fixed length and perpendicular to the equator when the other endpoint moves along the equator. Why is this curve not a “line”?

11.7. Al-Haytham’s attempted proof of the parallel postulate is fallacious because in non-Euclidean geometry two straight lines cannot be equidistant at all points. Thus in a non-Euclidean space the two rails of a railroad cannot both be straight lines. Assuming Newton’s laws of motion (an object that does not move in a straight line must be subject to some force), show that in a non-Euclidean universe one of the wheels in a pair of opposite wheels on a train must be subject to some unbalanced force at all times. [Note: The spherical earth that we live on happens to be non-Euclidean. Therefore the pairs of opposite wheels on a train cannot both be moving in a great circle on the earth’s surface.]

11.8. Prove that in any geometry, if a line passes through the midpoint of side AB of triangle ABC and is perpendicular to the perpendicular bisector of the side BC , then it also passes through the midpoint of AC . (This is easier than it looks: Consider the line that *does* pass through both midpoints, and show that it is perpendicular to the perpendicular bisector of BC ; then argue that there is only one line passing through the midpoint of BC that is perpendicular to the perpendicular bisector of BC .)

11.9. Use the previous result to prove, independently of the parallel postulate, that the line joining the midpoints of the lateral sides of a Thabit (Saccheri) quadrilateral bisects the diagonals.

CHAPTER 12

Modern Geometries

In geometry, as in number theory, the seventeenth century represents a break with the past. The two main reasons for the sudden surge of mathematical activity are the same in both cases: first, the availability of translations from the Arabic, which stimulated European mathematicians to try to recover and extend the fascinating results achieved by the ancient Greeks and medieval Muslims; second, the development of algebra and its evolution into a symbolic form in the Italian city-states during the sixteenth century. This development suggested new ways of thinking about old problems. The result was a variety of new forms of geometry that came about as a result of the calculus: analytic geometry, algebraic geometry, projective geometry, descriptive geometry, differential geometry, and topology.

1. Analytic and algebraic geometry

The creation of what we now know as analytic geometry had to wait for algebraic thinking about geometry (the type of thinking Pappus called *analytic*) to become a standard mode of thinking. No small contribution to this process was the creation of the modern notational conventions, many of which were due to François Viète (1540–1603) and Descartes. It was Descartes who started the very useful convention of using letters near the beginning of the alphabet for constants and data and those near the end of the alphabet for variables and unknowns. Viète's convention, which was followed by Fermat, had been to use consonants and vowels respectively for these purposes.

1.1. Fermat. Besides working in number theory, Fermat studied the works of Apollonius, including references by Pappus to lost works. This study inspired him to write a work on plane and solid loci, first published with his collected works in 1679. He used these terms in the sense of Pappus: A plane locus is one that can be constructed using straight lines and circles, and a solid locus is one that requires conic sections for its construction. He says in the introduction that he hopes to systematize what the ancients, known to him from Book 7 of Pappus' *Synagōgē*, had left haphazard. Pappus had written that the locus to more than six lines had hardly been touched. Thus, locus problems were the context in which Fermat invented analytic geometry.

Apart from his adherence to a dimensional uniformity that Descartes (finally!) eliminated, Fermat's analytic geometry looks much like what we are now familiar with. He stated its basic principle very clearly, asserting that the lines representing two unknown magnitudes should form an angle that would usually be assumed a right angle. He began with the equation of a straight line:¹ $Z^2 - DA = BE$. This equation looks strange to us because we automatically (following Descartes) tend

¹ Fermat actually wrote " Z pl. $- D$ in A æquetur B in E ."

to look at the Z as a variable and the A and E as constants, exactly the reverse of what Fermat intended. If we make the replacements $Z \mapsto c$, $D \mapsto a$, $A \mapsto x$, $E \mapsto y$, this equation becomes $c^2 - ax = by$, and now only the exponent looks strange, the result of Fermat's adherence to the Euclidean niceties of dimension.

Fermat illustrated the claim of Apollonius that a locus was determined by the condition that the sum of the pairwise products of lines from a variable point to given lines is given. His example was the case of two lines, where it is the familiar rectangular hyperbola that we have now seen used many times for various purposes. Fermat wrote its equation as $ae = z^2$. He showed that the graph of any quadratic equation in two variables is a conic section.

1.2. Descartes. Fermat's work on analytic geometry was not published in his lifetime, and therefore was less influential than it might have been. As a result, his contemporary René Descartes is remembered as the creator of analytic geometry, and we speak of "Cartesian" coordinates, even though Fermat was more explicit about their use.

René Descartes is remembered not only as one of the most original and creative modern mathematicians, but also as one of the leading voices in modern philosophy and science. Both his scientific work on optics and mechanics and his geometry formed part of his philosophy. Like Plato, he formed a grand project of integrating all of human knowledge into a single system. Also like Plato, he recognized the special place of mathematics in such a system. In his *Discourse on Method*, published at Leyden in 1637, he explained that logic, while it enabled a person to make correct judgments about inferences drawn through syllogisms, did not provide any actual knowledge about the world, what we would call empirical knowledge. In what was either a deadpan piece of sarcasm or a sincere tribute to Ramon Lull (mentioned above in Chapter 8), he said that in the art of "Lully" it enabled a person to speak fluently about matters on which he is entirely ignorant. He seems to have agreed with Plato that mathematical concepts are real objects, not mere logical relations among words, and that they are perceived directly by the mind. In his famous attempt at doubting everything, he had brought himself back from utter skepticism by deducing the principle that whatever he could clearly and distinctly perceive with his mind must be correct.

As Davis and Hersh (1986) have written, the *Discourse on Method* was the fruit of a decade and a half of hard work and thinking on Descartes' part, following a series of three vivid dreams on the night of November 10, 1619, when he was a 23-year-old soldier of fortune. The link between Descartes' philosophy and his mathematics lies precisely in the matter of "clear and distinct perception." For there seems to be no other area of thought in which human ideas are so clear and distinct. As Grabiner (1995, p. 84) says, when Descartes attacked, for example, a locus problem, the answer had to be "it is this curve, it has this equation, and it can be constructed in this way." Descartes' *Géométrie*, which contains his ideas on analytic geometry, was published as the last of three appendices to the *Discourse*.

What Descartes meant by "clear and distinct" ideas in mathematics is shown in a method of generating curves given in his *Géométrie* that appears mechanical, but can be stated in pure geometric language. A pair of lines intersecting at a fixed point Y coincide initially (Fig. 1). The point A remains fixed on the horizontal line. As the oblique line rotates about Y , the point B , which remains fixed on it, describes a circle. The tangent at B intersects the horizontal line at C , and

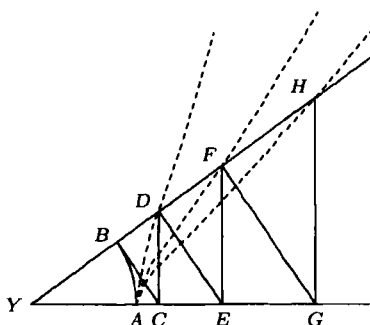


FIGURE 1. Descartes' linkage for generating curves. The curve $x^{4n} = a^2(x^2 + y^2)^{2n-1}$ is shown for $n = 0, 1, 2, 3$.

the point on the oblique line directly above C is D . The line perpendicular to the oblique line at D intersects the horizontal line at E , from which a vertical line intersects the oblique line at F , and so forth in a zigzag pattern. Descartes imagined a mechanical linkage that could actually draw these curves.

Descartes regarded determinate curves of this sort, depending on one parameter, as we would say, as legitimate to use in geometry. He offered the opinion that the opposition to "mechanical" curves by ancient Greek mathematicians arose because the curves they knew about—he mentioned the spiral of Archimedes and the quadratrix—were indeterminate. In the case of the spiral of Archimedes, which is generated by a point moving at constant linear velocity along a line that is rotating with constant angular velocity, the indeterminacy arises because the two velocities need to be coordinated with infinite precision. For the quadratrix, the same problem arises, as the ratio of the velocity of a rotating line and that of a translating line needs to be known with infinite precision.

Descartes' *Géométrie* resembles a modern textbook of analytic geometry less than does Fermat's *Introduction to Plane and Solid Loci*. He does not routinely use a system of "Cartesian" coordinates, as one might expect from the name. But he does remove the dimensional difficulties that had complicated geometric arguments since Euclid's cumbersome definition of a composite ratio.

[U]nity can always be understood, even when there are too many or too few dimensions; thus, if it be required to extract the cube root of $a^2b^2 - b$, we must consider the quantity a^2b^2 divided once by unity, and the quantity b multiplied twice by unity. [Smith and Latham, 1954, p. 6]

Here Descartes is explaining that all four arithmetic operations can be performed on *lines* and yield *lines* as a result. He illustrated the product and square root by the diagrams in Fig. 2, where $AB = 1$ on the left and $FG = 1$ on the right.

Descartes went a step further than Oresme in eliminating dimensional considerations, and he went a step further than Pappus in his classification of locus problems. Having translated these problems into the language of algebra, he realized that the three- and four-line locus problems always led to polynomial equations of degree at most 2 in x and y , and conversely, any equation of degree 2 or less represented a three- or four-line locus. He asserted with confidence that he had solved

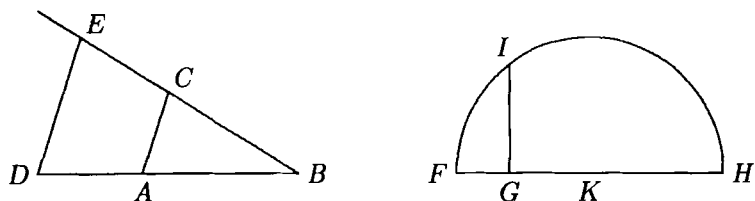


FIGURE 2. Left: $AB = 1$, so that $BE = BC \cdot BD$. Right: $FG = 1$, so that $GI = \sqrt{GH}$.

the problem that Pappus reported unsolved in his day. It was in this context that he formulated the idea of using two intersecting lines as a frame of reference, saying that

since so many lines are confusing, I may simplify matters by considering one of the given lines and one of those to be drawn... as the principal lines, to which I shall try to refer all the others. [Smith and Latham, 1954, p. 29]

The idea of using two coordinate lines is psychologically very close to the linkages illustrated in Fig. 1. In terms of Fig. 3, Descartes took one of the fixed lines as a horizontal axis AB , since a line was to be drawn from point C on the locus making a fixed angle θ with AB . He thought of this line as sliding along AB and intersecting it at point B , and he denoted the variable length AB by x . Then since C needed to slide along this moving line so as to keep the proportions demanded by the conditions of the locus problem, he denoted the distance CB by y . All the lines were fixed except CB , which moved parallel to itself, causing x to vary, while on it y adjusted to the conditions of the problem. For each of the other fixed lines, say AR , the angles ψ , θ , and φ will all be given, ψ by the position of the fixed lines AB and AR , and the other two by the conditions prescribed in the problem. Since these three angles determine the shape of the triangles ADR and BCD , they determine the ratios of any pair of sides in these triangles through the law of sines, and hence all sides can be expressed in terms of constants and the two lengths x and y . If the set of $2n$ lines is divided into two sets of n as the $2n$ -line locus problem requires, the conditions of the problem can be stated as an equation of the form

$$p(x, y) = q(x, y),$$

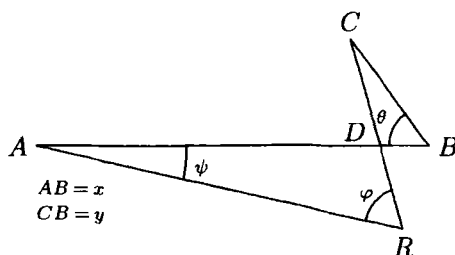
where p and q are of degree at most n in each variable. The analysis was mostly "clear and distinct."

Descartes argued that the locus could be considered known if one could locate as many points on it as desired.² He next pointed out that in order to locate points on the locus one could assign values to either variable x and y , then compute the value of the other by solving the equation.³

Everyone who has studied analytic geometry in school must have been struck at the beginning by how much clearer and easier it was to use than the synthetic geometry of Euclid. That aspect of the subject is nicely captured in the words the poet Paul Valéry (1871–1945) applied to Descartes' philosophical method in

² The validity of this claim is somewhat less than "clear and distinct."

³ This claim also involves a great deal of hope, since equations of degree higher than 4 were unknown territory in his day.

FIGURE 3. Descartes' analysis of the n -line locus problem.

general: "the most brilliant victory ever achieved by a man whose genius was applied to reducing the need for genius" (quoted by Davis and Hersh, 1986, p. 7).

This point was ignored by Newton in a rather ungenerous exhibition of his own remarkable mathematical talent (Whiteside, 1967, Vol. IV, pp. 275–283). Newton said that Descartes "makes a great show" about his solution of the three- and four-line locus problems, "as if he had achieved something so earnestly sought after by the ancients." He also expressed a distaste for Descartes' use of symbolic algebra to solve this problem (a distaste that would be echoed by other mathematicians), saying that if this algebra were written out in words, it "would prove to be so tedious and entangled as to provoke nausea." One is inclined to say, on Descartes' behalf, "Precisely! That's why it's better to use algebraic symbolism and avoid the tedium, confusion, and nausea."

1.3. Newton's classification of curves. Like Descartes, Newton made a classification of curves according to the degree of the equations that represent them, or rather, according to the maximal number of points in which they could intersect a straight line. As Descartes had argued for the use of any curves that could be generated by one parameter, excluding spirals and the quadratrix because they required two independent motions to be coordinated, Newton likewise argued that geometers should either confine themselves to conic sections or else allow any curve having a clear description. In his *Universal Arithmetick*, he mentioned in particular the trochoid,⁴ which makes it possible to divide an angle into any number of equal parts, as a useful curve that is simple to describe.

1.4. Algebraic geometry. As we have just seen, Descartes began the subject of algebraic geometry with his classification of algebraic curves into genera, and Newton gave an alternative classification of curves also, based on algebra, although he included some curves that we would call transcendental, curves that could intersect a line in infinitely many points. The general study of algebraic curves $p(x, y) = 0$, where $p(x, y)$ is a polynomial in two variables, began with Colin Maclaurin (1698–1746), who in his *Geometria organica* of 1720 remarked that a cubic curve was not uniquely determined by nine points, even though nine points apparently suffice to determine the coefficients of any polynomial $p(x, y)$ of degree 3, up to proportionality and hence determine a unique curve $p(x, y) = 0$. Actually, however, two *distinct*

⁴ A trochoid is the locus of a point rigidly attached to a rolling wheel. If the point lies between the rim and the center, the trochoid is called a *curtate cycloid*. If the point lies outside the rim, the trochoid is a *prolate cycloid*. If the point is right on the rim, the trochoid is called a *cycloid*. The names come from the Greek words *trokhós* (*wheel*) and *kýklos* (*circle*).

cubic curves generally intersect in nine points, so that *some* sets of nine points do not determine the curve uniquely (see Problem 12.7). This fact was later (1748) noted by Euler as well, and finally, by Gabriel Cramér (1704–1752), who also noted Maclaurin's priority in the discovery that a curve of degree m and a curve of degree n meet generally in mn points. This interesting fact is called *Cramér's paradox* after Cramér published it in a 1750 textbook on algebraic curves. Although he correctly explained why more than one curve of degree n can sometimes be made to pass through $n(n+3)/2$ points—because the equations for determining the coefficients from the coordinates of the points might not be independent—he noted that in that case there were actually infinitely many such curves. That, he said, was a real paradox. Incidentally, it was in connection with the determination of the coefficients of an algebraic curve through given points that Cramér stated Cramér's rule for solving a system of linear equations by determinants.⁵

2. Projective and descriptive geometry

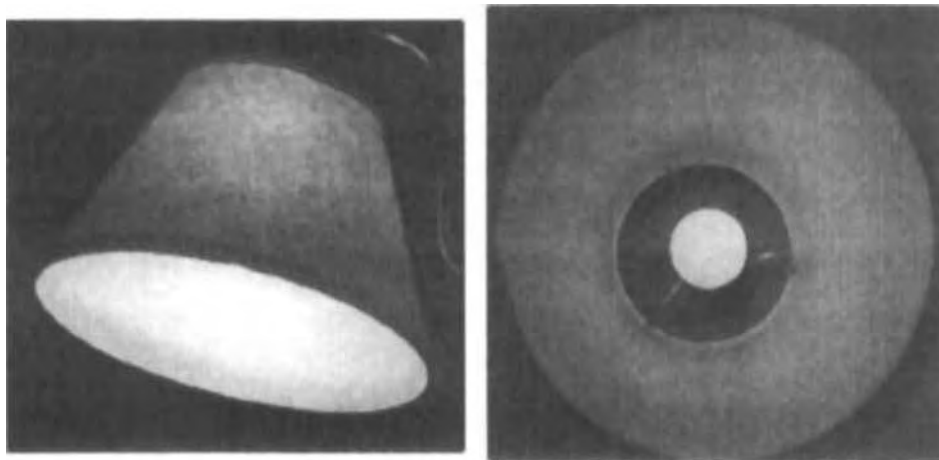
It is said that Euclid's geometry is tactile rather than visual, since the theorems tell you what you can measure and feel with your hands, not what your eye sees. It is a commonplace that a circle seen from any position except a point on the line through its center perpendicular to its plane appears to be an ellipse. If figures did not distort in this way when seen in perspective, we would have a very difficult time navigating through the world. We are so accustomed to adjusting our judgments of what we see that we usually recognize a circle automatically when we see it, even from an angle. The distortion is an essential element of our perception of depth. Artists, especially those of the Italian renaissance, used these principles to create paintings that were astoundingly realistic. As Leonardo da Vinci (1452–1519) said, “the primary task of a painter is to make a flat plane look like a body seen in relief projecting out of it.” Many records of the principles by which this effect was achieved have survived, including treatises of Leonardo himself and a very famous painter's manual of Albrecht Dürer (1471–1528), first published in 1525. Over a period of several centuries these principles gave rise to the subject now known as projective geometry.

2.1. Projective properties. Projective geometry studies the mathematical relations among figures that remain constant in perspective. Among these things are points and lines, the number of intersections of lines and circles, and consequently also such things as parallelism and tangency, but not things that depend on shape, such as angles or circles.

A less obvious property that is preserved is what is now called the *cross-ratio* of four points on a line.⁶ If A , B , C , and D are four points on a line, with B and

⁵ As mentioned in Chapter 8, the solution of linear equations by determinants had been known to Seki Kōwa and Leibniz. Thus, Cramér has two mathematical concepts named after him, and in both cases he was the third person to make the discovery.

⁶ Although this ratio has been used for centuries, the name it now bears in English seems to go back only to an 1869 treatise on dynamics by William Kingdon Clifford (1845–1879). Before that it was called the *anharmonic ratio*, a phrase translated from an 1837 French treatise by Michel Chasles (1809–1880). This information came from the website on the history of mathematical terms maintained by Jeff Miller of Gulf High School in New Port Richey, Florida. The url of the website is <http://members.aol.com/jeff570/mathword.html>.



A circle seen in perspective is an ellipse.

C both between A and D and C between B and D , their cross-ratio is

$$(A, B, C, D) = \frac{AC \cdot BD}{AD \cdot BC}.$$

It is not difficult to show, for example, that if the rays PA , PB , PC , and PD from a point P intersect a second line in points A' , B' , C' , and D' , the cross-ratio of these new points is the same as that of the original four points. Coolidge (1940, p. 88) speculated that Euclid may have known about the cross-ratio, and he asserted that the early second-century mathematician Menelaus did know about it.

Some theorems that might appear difficult to prove from the standard Euclidean techniques of proportion and congruence can be quite easy when looked at “in perspective,” so to speak. For example, it is easy to prove that if two tangents to a circle from points A and C meet at a point P , then the line from P to the midpoint of the chord AC meets the circle in a point (namely the midpoint of the arc \widehat{AC}) at which the tangent to the circle is parallel to the chord AC . To prove that same theorem for an ellipse using analytic geometry is a very tedious computation. However, remembering that the ellipse was obtained as the intersection of a cone with a plane oblique to its base, one has only to note that projection preserves tangency, intersections, parallelism (usually), and midpoints. Then, projecting the cone and all the lines into the base plane yields the result immediately, as shown in Fig. 4.⁷ Similarly, it could be shown that the bisector of a chord from the point of intersection of the tangents at the endpoints of the chord passes through the center of the ellipse.

2.2. The Renaissance artists. The revival of interest in ancient culture in general during the Renaissance naturally carried with it an interest in geometry. The famous artist Piero della Francesca (1410?–1492) was inspired by the writings of Leonardo of Pisa and others to write treatises on arithmetic and the five regular solids. The scholar Luca Pacioli (1445–1517), who was influenced by Piero della Francesca and was a friend of Leonardo da Vinci, published a comprehensive

⁷ Of course, in the figure the “circle” in the base is really an ellipse because it has been projected onto the page.

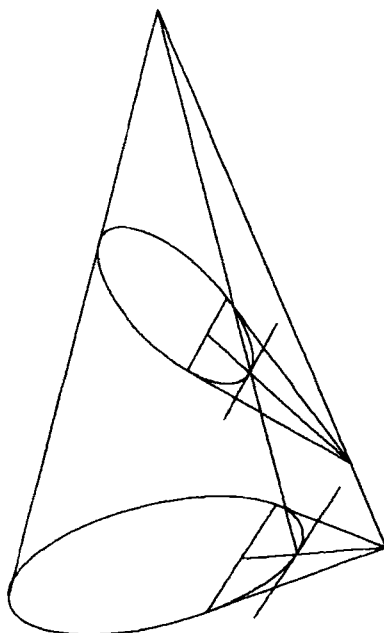


FIGURE 4. Central projections preserve tangency, midpoints, and (usually) parallelism.

treatise on arithmetic and geometry in 1494, and a second book, *De divina proportione*, in 1509. He gave the name *divine proportion* to what is now called the *Golden Section*, the division of a line into mean and extreme ratios. Interest in the five regular solids branched out into an interest in semiregular solids. Leonardo da Vinci designed wooden models of these, which were depicted in Pacioli's treatise.

The regular and semiregular solids formed an important part of Dürer's manual for painters, published in 1525. He showed how to cut out a paper model of a truncated icosahedron, which consists of 12 pentagons and 20 hexagons (Fig. 5). The solid, although not the name, has become very familiar to modern people through its application in athletics and organic chemistry.

A geometric description of perspective was given by Leon Battista Alberti (1404–1472) in a treatise entitled *Della pittura*, published posthumously in 1511. If the eye is at fixed height above a horizontal plane, parallel horizontal lines in that plane receding from the imagined point where the eye is located can be drawn as rays emanating from a point (the vanishing point) at the same height above the plane, giving the illusion that the vanishing point is infinitely distant. The application to art is obvious: Since the canvas can be thought of as a window through which the scene is viewed, if you want to draw parallel horizontal lines as they would appear through a window, you must draw them as if they all converged on the vanishing point. Thus, a family of lines having a common property (passing through the vanishing point) projects to a family having a different common property (being parallel to one another). Obviously, lines remain lines under such a projection. However, perpendicular lines will not remain perpendicular, nor will circles remain circles.

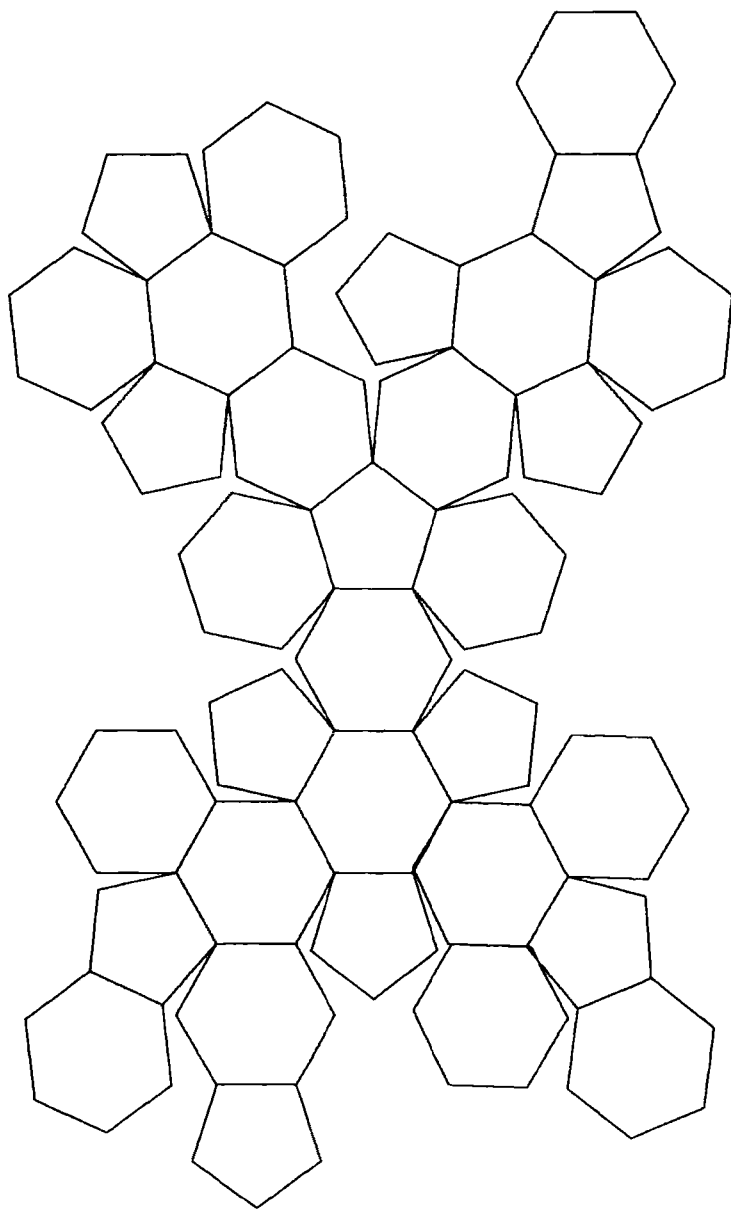
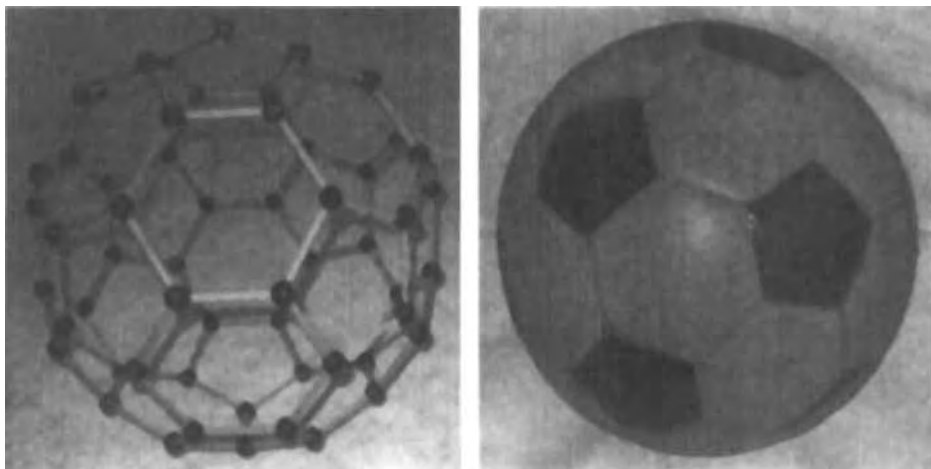


FIGURE 5. Dürer's paper model of a truncated icosahedron.

In those days before photography and computers, the mechanical aspects of drawing according to Alberti's rules apparently did not disturb artists. Dürer, in particular, seemed to enjoy thinking up mechanical ways of producing technical perfection. One of his devices is shown below. Although the device seems a very



Two modern applications of the truncated icosahedron: a molecule of buckminsterfullerene ("buckyball"); a soccer ball.

strange and inefficient way of painting, it does illustrate the use of projection very vividly, even if it was only a "thought experiment."⁸

2.3. Girard Desargues. The mathematical development of the theory of projection began with the work of Girard Desargues (1593–1662). In 1636, one year before the publication of Descartes' *Géométrie*, Desargues published a pamphlet with the ponderous title *An Example of One of the General Methods of S.G.D.L.*⁹ *Applied to the Practice of Perspective Without the Use of Any Third Point, Whether of Distance or Any Other Kind, Lying Outside the Work Area.* The reference to a "third point" was aimed at the primary disadvantage of Alberti's rules, the need to use a point not on the canvas in order to get the perspective correct. Three years later he produced a *Rough Draft of an Essay on the Consequences of Intersecting a Cone with a Plane*. In both works, written in French rather than the more customary Latin, he took advantage of the vernacular to invent new names, not only for the conic sections,¹⁰ as Dürer had done, but also for a large number of concepts that called attention to particular aspects of the distribution and proportions of points and lines. He was particularly fond of botanical names,¹¹ and included *tree*, *trunk*, *branch*, *shoot*, and *stem*, among many other neologisms. Although the new language might seem distracting, using standard terms for what he had in mind would have been misleading, since the theory he was constructing unified concepts that had been distinct before. For example, he realized that a cylinder could be regarded as a limiting case of a cone, and so he gave the name *scroll* to the class consisting of both surfaces. Desargues had very little need to refer to any specific conic section; his theorems applied to all of them equally. As he said (Field and Gray, 1987, p. 102—I have changed their *roll* to *scroll*):

⁸ According to Strauss (1977, p. 31), painters of Dürer's time who actually tried to build such devices found them quite impracticable.

⁹ Sieur Girard Desargues Lyonnois.

¹⁰ He gave the standard names, but suggested *deficit*, *equalation*, and *exceedence* as alternatives.

¹¹ Ivins (1947, cited by Field and Gray, 1987, p. 62) suggested that these names were inspired by similar names in Alberti's treatise.



One of Dürer's devices for producing an accurate painting. The artist's assistant at the left holds a needle at a particular point on the lute being painted, while the artist sticks a pair of crosshairs on the frame to mark the exact point where the thread passes through the window. The needle and thread are then to be removed, the door holding the canvas closed, and the spot where the crosshairs meet marked on the canvas. © Corbis Images (No. SF1906).

The most remarkable properties of the sections of a scroll are common to all types, and the names *Ellipse*, *Parabola*, and *Hyperbola* have been given them only on account of matters extraneous to them and to their nature.

Desargues was among the first to regard lines as infinitely long, in the modern way. In fact, he opens his treatise by saying that he will consider both the infinitely large and the infinitely small in his work, and he says firmly that "in this work every straight line is, if necessary, taken to be produced to infinity in both directions." He also had the important insight that a family of parallel lines and a family of lines with a common point of intersection have similar properties. He said that lines belonged to the same *order*¹² if either they all intersected at a common point or were all mutually parallel. This term was introduced "[to] indicate that in the one case as well as in the other, it is *as if* they all converged to the same place" [emphasis added].

¹² Now called a *pencil* or *sheaf*.

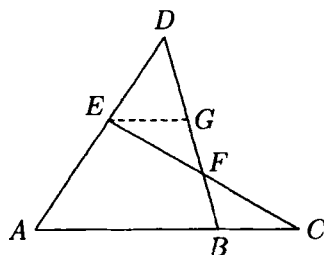


FIGURE 6. Menelaus' theorem for a plane triangle.

Although Desargues' terminology is very difficult to follow, his *Rough Draft* contained some elegant theorems about points on conics. Two significant results are the following:¹³

First: *If four lines in a plane intersect two at a time, and the points of intersection on the first line are A, B, and C, with B between A and C, and the lines through A and B intersect in the point D, those through A and C in E and those through B and C in F, then*

$$(1) \quad \frac{BD}{BF} = \frac{AD}{AE} \cdot \frac{CE}{CF}.$$

The situation here was described by Pappus, and the result is also known as Menelaus' theorem. The proof is easily achieved by drawing the line through E parallel to AB, meeting BD in a point G, then using the similarity of triangles EGF and CBF and of triangles DEG and DAB, as in Fig. 6. From Eq. 1 it is easy to deduce that $BD \cdot AE \cdot CF = BF \cdot AD \cdot CE$. Klein (1926, p. 80) attributes this form of the theorem to Lazare Carnot (1753–1823).

Second: *The converse of this statement is also true, and can be interpreted as stating that three points lie on a line.* That is, if ADB is a triangle, and E and F are points on AD and BD respectively such that $AD : AE < BD : BF$, then the line through E and F meets the extension of AB on the side of B in a point C, which is characterized as the only point on the line EF satisfying Eq. 1.

In 1648 the engraver Abraham Bosse (1602–1676), who was an enthusiastic supporter of Desargues' new ideas, published *La Perspective de Mr Desargues*, in which he reworked these ideas in detail. Near the end of the book he published the theorem that is now known as Desargues' theorem. Like Desargues' work, Bosse's statement of the theorem is a tangled mess involving ten points denoted by four uppercase letters and six lowercase letters. The points lie on nine different lines. When suitably clarified, the theorem states that if the lines joining the three pairs of vertices from two different triangles intersect in a common point, the pairs of lines containing the corresponding sides of these triangles meet in three points all on the same line. This result is easy to establish if the triangles lie in different planes, since the three points must lie on the line of intersection of the two planes containing the triangles, as shown in Fig. 7.

For two triangles in the same plane, the theorem, illustrated in Fig. 8, was proved by Bosse by applying Menelaus' theorem to the three sets of collinear points

¹³ To keep the reader's eye from getting *too* tangled up, we shall use standard letters in the statement and figure rather than Desargues' weird mixture of uppercase and lowercase letters and numbers, which almost seems to anticipate the finest principles of computer password selection.

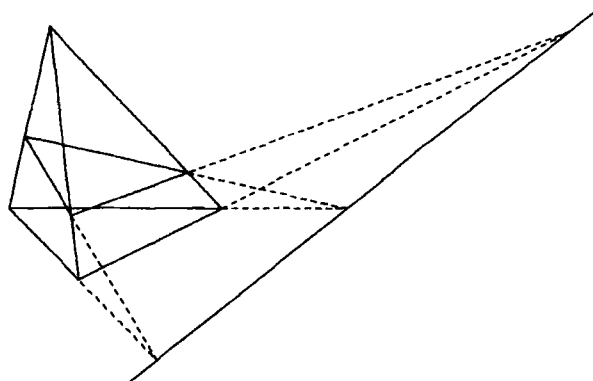


FIGURE 7. Desargues' theorem for triangles lying in different planes.

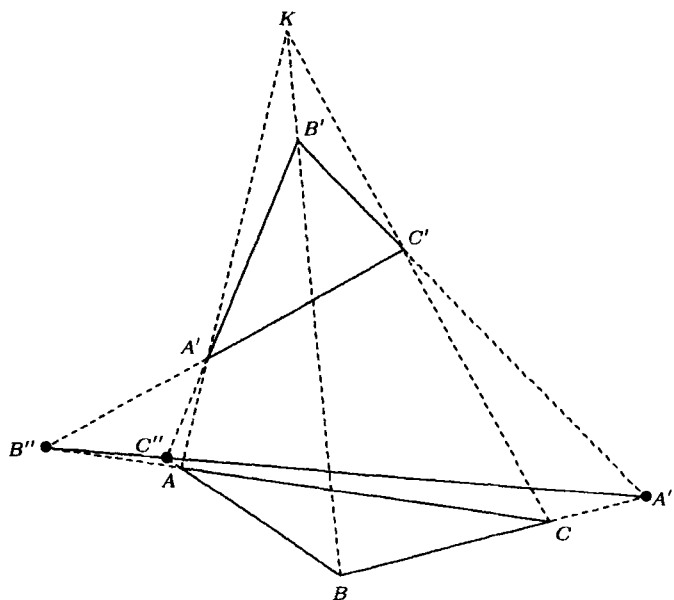


FIGURE 8. Desargues' theorem for two triangles in the same plane.

$\{A'', C, B\}$, $\{B'', A, C\}$, and $\{C'', A, B\}$, with K as the third vertex of the triangle whose base ends in the second and third points in all three cases. (There is no other conceivable way to proceed, so that in a sense the proof is a mere computation.) When the ratios $AK : AA'$, $BK : BB'$, and $CK : CC'$ are eliminated from the three resulting equations, the result can be written as the equation

$$\frac{C''B'}{C''A'} = \frac{A''B'}{A''C'} \cdot \frac{B''C'}{B''A'}.$$

Having received a copy of this work from Marin Mersenne, Descartes took the word *draft* literally and regarded it as a proposal to write a treatise—which it may have been—such as a modern author would address to a publisher, and a publisher would send to an expert for review. He wrote to Desargues to express his opinion of “what I can conjecture of the *Treatise on Conic Sections*, of which [Mersenne] sent me the *Draft*.” Descartes’ “review” of the work contained the kind of advice reviewers still give: that the author should decide more definitely who the intended audience was. As he said, if Desargues was aiming to present new ideas to scholars, there was no need to invent new terms for familiar concepts. On the other hand, if the book was aimed at the general public, it would need to be very thick, since everything would have to be explained in great detail (Field and Gray, 1987, p. 176).

2.4. Blaise Pascal. Desargues’ work was read by a teenage boy named Blaise Pascal (1623–1662), who was to become famous for his mathematical work and renowned for his *Pensées* (*Meditations*), which are still read by many people today for inspiration. He began working on the project of writing his own treatise on conics. Being very young, he was humble and merely sketched what he planned to do, saying that his mistrust of his own abilities inclined him to submit the proposal to experts, and “if someone thinks the subject worth pursuing, we shall try to carry it out to the extent that God gives us the strength.” Pascal admired Desargues’ work very much, saying that he owed “what little I have discovered to his writings” and would imitate Desargues’ methods, which he considered especially important because they treated conic sections without introducing the extraneous axial section of the cone. He did indeed use much of Desargues’ notation for points and lines, including the word *order* for a family of concurrent lines. His work, like that of Desargues, remained only a draft, although Struik (1986, p. 165) reports that Pascal did work on this project and that Leibniz saw a manuscript of it—not the rough draft, apparently—in 1676. All that has been preserved, however, is the rough draft. That draft contains several results in the spirit of Desargues, one of which, called by Pascal a “third lemma,” is still known as Pascal’s theorem. Referring to Fig. 9, in which four lines MK , MV , SK , and SV are drawn and then a conic is passed through K and V meeting these four lines in four other points P , O , N , and Q respectively, Pascal asserted that the lines PQ , NO , and MS would be concurrent (belong to the same *order*).

2.5. Newton’s degree-preserving mappings. Newton also made contributions to projective geometry, in a way that related it to Descartes’ analytic geometry and to algebraic geometry. He described the mapping shown in Fig. 10 (Whiteside, 1967, Vol. VI, p. 269). In that figure the parallel lines BL and AO and the points A , B , and O are fixed from the outset, and the angle θ is specified in advance. Thus the distances h and Δ and the angles φ and θ are given before the mapping is defined. Then, to map the figure GHI to its image ghi , first project each point G parallel to BL so as to meet the extension of AB at a point D . Next, draw the line OD meeting BL in point d . Finally, from d along the line making angle θ with BL , choose the image point g so that $gd : Od :: GD : OD$. The original point, according to Newton, had coordinates (BD, DG) and its image the coordinates (Bd, dg) . Thus, if we let $x = BD$ and $y = DG$, the coordinate transformation in

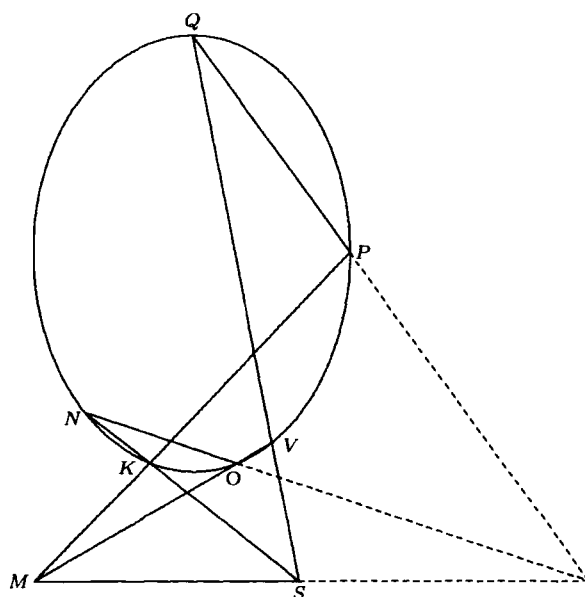


FIGURE 9. Pascal's theorem (third lemma).

the two directions can be described as

$$\begin{aligned}(x, y) &\mapsto (\xi, \eta) = \left(\frac{\Delta x \sin \varphi}{h + x \sin \varphi}, \frac{hy}{h + x \sin \varphi} \right), \\ (\xi, \eta) &\mapsto \left(\frac{h\xi}{(\Delta - \xi) \sin \varphi}, \frac{\Delta \eta}{\Delta - \xi} \right).\end{aligned}$$

Newton noted that this kind of projection preserves the degree of an equation. Hence a conic section will remain a conic section, a cubic curve will remain a cubic curve, and so on, under such a mapping. In fact, if a polynomial equation $p(x, y) = 0$ is given whose highest-degree term is $x^m y^n$, then every term $x^p y^q$, when expressed in terms of ξ and η , will be a multiple of $\xi^p \eta^q / (\Delta - \xi)^{p+q}$, so that if the entire equation is converted to the new coordinates and then multiplied by $(\Delta - \xi)^{m+n}$, this term will become $\xi^p \eta^q (\Delta - \xi)^{m+n-p-q}$, which will be of degree $m+n$. Thus the degree of an equation does not change under Newton's mapping. These mappings are special cases of the transformations known as *fractional-linear* or *Möbius* transformations, after August Ferdinand Möbius (1790–1868), who developed them more fully. They play a vital role in algebraic geometry and complex analysis, being the only one-to-one analytic mappings of the extended complex plane onto itself. According to Coolidge (1940, p. 269), it was Edward Waring (1736–1798) who first remarked, in 1762, that fractional-linear transformations were the most general degree-preserving transformations.

2.6. Charles Brianchon. Pascal's work on the projective properties of conics was extended by Charles Julien Brianchon (1785–1864), who was also only a teenager when he proved what is now recognized as the dual of Pascal's theorem: *The pairs*

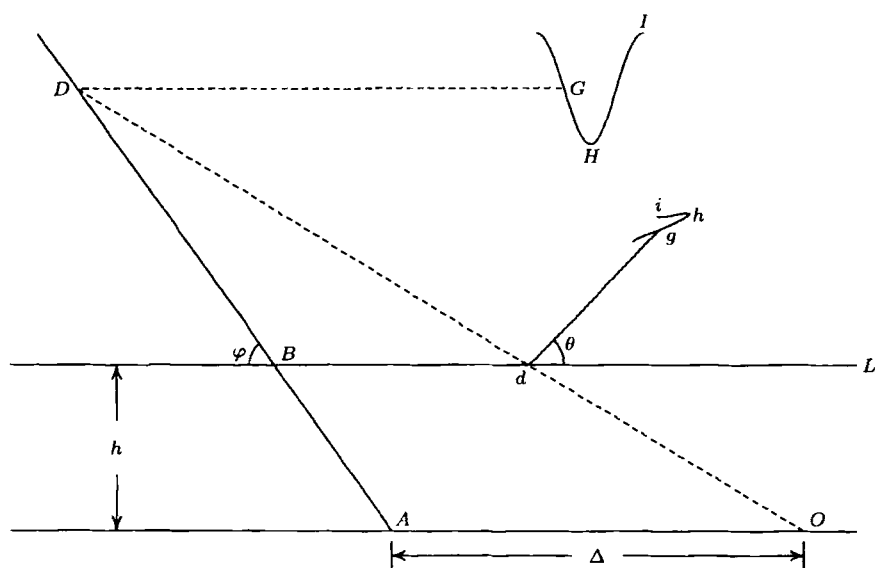


FIGURE 10. Newton's degree-preserving projection.

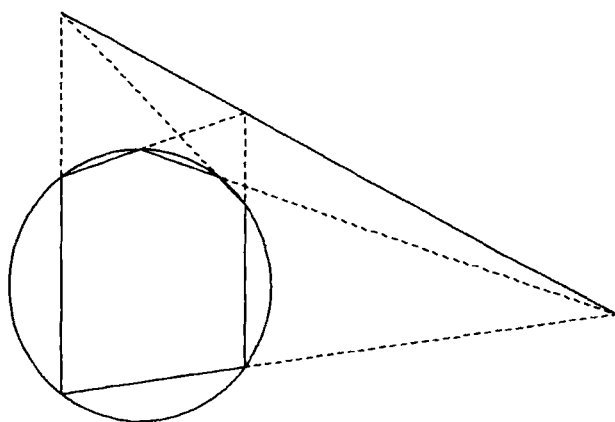


FIGURE 11. Brianchon's theorem for a circle.

of opposite sides of a hexagon inscribed in a conic meet in three collinear points. The case of a circle is illustrated in Fig. 11.

2.7. Monge and his school. After a century of relative neglect, projective geometry revived at the École Polytechnique under the students of Gaspard Monge (1746–1818), who was a master of the application of calculus to geometry. Klein (1926, pp. 77–78) described his school as distinguished by “the liveliest spatial intuition combined in the most natural way possible with analytic operations.” Klein went on to say that he taught his students to make physical models, “not to make up for the deficiencies of their intuition but to develop an already clear and lively intuition.” As a military engineer, Monge had used his knowledge of geometry to design fortifications. His work in this area was highly esteemed by his superiors

and declared a military secret. He wrote a book on descriptive geometry and one on the applications of analysis to geometry, whose influence appeared in the work of his students. Klein says of the second book that it "reads like a novel." In this book, Monge analyzed quadric surfaces with extreme thoroughness.

Monge is regarded as the founder of descriptive geometry, which is based on the same principles of perspective as projective geometry but more concerned with the mechanics of representing three-dimensional objects properly in two dimensions and the principles of interpreting such representations. Monge himself described the subject as the science of giving a complete description in two dimensions of those three-dimensional objects that can be defined geometrically. As such, it continues to be taught today under other names, such as mechanical drawing; it is the most useful form of geometry for engineers.

Monge's greatest student (according to Klein) was Jean-Victor Poncelet (1788–1867). He participated as a military engineer in Napoleon's invasion of Russia in 1812, was wounded, and spent a year in a Russian prison, where he busied himself with what he had learned from Monge. Returning to France, he published his *Treatise on the Projective Properties of Figures* in 1822, the founding document of modern projective geometry. Its connection with its historical roots in the work of Desargues shows in the first chapter, where Poncelet says he will be using the word *projective* in the same sense as the word *perspective*. In Chapter 3 he introduces the idea that all points at infinity in a plane can be regarded as belonging to a single line at infinity.¹⁴ These concepts brought out fully the duality between points and lines in a plane and between points and planes in three-dimensional space, so that interchanging these words in a theorem of projective geometry results in another theorem. The theorems of Pascal and Brianchon, for example, are dual to each other.

2.8. Jacob Steiner. The increasing algebraization of geometry was opposed by the Swiss mathematician Jacob Steiner (1796–1863), described by Klein (1926, pp. 126–127) as "the only example known to me... of the development of mathematical abilities after maturity." Steiner had been a farmer up to the age of 17, when he entered the school of the Swiss educational reformer Johann Heinrich Pestalozzi (1746–1827), whose influence was widespread, extending through the philosopher-psychologist Johann Friedrich Herbart (1776–1841) down to Riemann, as will be explained in the next section.¹⁵ Steiner was a peculiar character in the history of mathematics, who when his own originality was in decline, adopted the ideas of others as his own without acknowledgement (see Klein, 1926, p. 128). But in his best years, around 1830, he had the brilliant idea of building space using higher-dimensional objects such as lines and planes instead of points, recognizing that these objects were projectively invariant. He sought to restore the ancient Greek "synthetic" approach to geometry, which was independent of numbers and the concept of length. To this end, in his 1832 work on geometric figures he considered a family of mappings of one plane on another that resembles somewhat Newton's projection. Klein (1926, p. 129) found nothing materially new in this work, but admired the systematization that it contained. The Steiner principle of successively

¹⁴ Field and Gray (1987, p. 185) point out that Johannes Kepler (1571–1630) had introduced points at infinity in a 1604 work on conic sections, so that a parabola would have two foci.

¹⁵ Klein (1926, pp. 127–128), has nothing good to say about the more extreme recommendations of these men, calling these recommendations "pedagogical monstrosities."

building more and more intricate figures by allowing simpler ones to combine geometrically was novel and had its uses, but according to Klein, encompassed only one part of geometry.

2.9. August Ferdinand Möbius. Projective geometry was enhanced through the barycentric calculus invented by August Ferdinand Möbius (1790–1868) and expounded in a long treatise in 1827. This work contained a number of very useful innovations. Möbius' use of barycentric coordinates to specify the location of a point anticipated vector methods by some 20 years, and proved its value in many parts of geometry. He used his barycentric coordinates to classify plane figures in new ways. As he explained in Chapter 3 of the second section of his barycentric calculus (Baltzer, 1885, pp. 177–194), if the vertices of a triangle were specified as A, B, C , and one considered all the points that could be written as $aA + bB + cC$, with the lengths of the sides and the proportions of the coefficients $a : b : c$ given, all such figures would be congruent (he used the phrase “equal and similar”). If one specified only the proportions of the sides instead of their lengths, all such figures would be similar. If one specified only the proportions of the coefficients, the figures would be in an *affine* relationship, a word still used to denote a linear transformation followed by a translation in a vector space. Finally, he introduced the relation of equality (in area).

Cauchy, then at the height of his powers, reviewed Möbius' work¹⁶ on the barycentric calculus. In his review, as reported by Baltzer (1885, pp. xi–xii), he was cautious at first, saying that the work was “a different method of analytic geometry whose foundation is certainly not so simple; only a deeper study can enable us to determine whether the advantages of this method will repay the difficulties.” After reporting on the new classification of figures in Part 2, he commented:

One must be very confident of taking a large step forward in science to burden it with so much new terminology and to demand that your readers follow you in investigations presented to them in such a strange manner.

Finally, after reporting some of the results from Part 3, he concluded that, “It seems that the author of the barycentric calculus is not familiar with the general theory of duality between the properties of systems of points and lines established by M. Gergonne.” This comment is difficult to explain on the assumption that Cauchy had actually read Chapters 4 and 5 of Part 3, since this duality (*gegenseitiges Entsprechen*) was part of the title of both chapters; but perhaps Cauchy was alluding to ideas in Gergonne's papers not found in the work of Möbius. Chapters 4 and 5 contain some of the most interesting results in the work. Chapter 4, for example, discusses conic sections and uses the barycentric calculus to prove that two distinct parabolas can be drawn through four coplanar points, provided none of them lies inside the triangle formed by the other three.

¹⁶ It might appear that Cauchy was able to read German, not a common accomplishment for French mathematicians in the 1820s, when the vast majority of mathematical papers of significance were written in French. But perhaps he read a French or Latin version of the work.

Möbius is best remembered for two concepts, the Möbius transformation, and the Möbius band. A Möbius transformation, by which we now understand a mapping of the complex plane into itself, $z \mapsto w$, of the form

$$w = \frac{az + b}{cz + d}, \quad ad - bc \neq 0,$$

can be found in his 1829 paper on metric relations in line geometry. He gave such transformations with real coefficients in terms of the two coordinates (x, y) , the real and imaginary parts of what we now write as the complex number z , and showed that they were the most general one-to-one transformations that preserve collinearity. The Möbius band is discussed in Section 4 below.

2.10. Julius Plücker. A number of excellent German, Swiss, and Italian geometers arose in the nineteenth century. Their work cannot be classified as purely projective geometry, since it also relates to algebraic geometry. As an example, we take Julius Plücker (1801–1868), who was a professor at the University of Bonn for the last 30 years of his life. Plücker himself remembered (Coolidge, 1940, p. 144) that when young he had discovered a theorem in Euclidean geometry: The three lines containing the common chords of pairs of three intersecting circles are all concurrent. Plücker's proof of this theorem is simplicity itself. Suppose that the equations of the three circles are $A = 0$, $B = 0$, $C = 0$, where each equation contains $x^2 + y^2$ plus linear terms. By subtracting these equations in pairs, we get the quadratic terms to drop out, leaving the equations of the three lines containing the three common chords: $A - B = 0$, $A - C = 0$, $B - C = 0$. But it is manifest that any two of these equations imply the third, so that the point of intersection of any two also lies on the third line.

Plücker's student Felix Klein (1926, p. 122) described a more sophisticated specimen of this same kind of reasoning by Plücker to prove Brianchon's theorem¹⁷ that the opposite sides of a hexagon inscribed in a conic, when extended, intersect in three collinear points. The proof goes as follows: The problem involves two sets, each containing three lines, six of whose nine pairwise intersections lie on a conic section. The conic section has an equation of the form $q(x, y) = 0$, where $q(x, y)$ is quadratic in both x and y . Represent each line by a linear polynomial of the form $a_jx + b_jy + c_j$, the j th line being the set of (x, y) where this polynomial equals zero, and assume that the lines are numbered in clockwise order around the hexagon. Form the polynomial

$$s(x, y) = (a_1x + b_1y + c_1)(a_3x + b_3y + c_3)(a_5x + b_5y + c_5) \\ - \mu(a_2x + b_2y + c_2)(a_4x + b_4y + c_4)(a_6x + b_6y + c_6)$$

with the parameter μ to be chosen later. This polynomial vanishes at all nine intersections of the lines. Line 1, for example, meets lines 2 and 6 inside the conic and line 4 outside it.¹⁸

Now, when y is eliminated from the equations $q(x, y) = 0$ and $s(x, y) = 0$, the result is an equation $t(x) = 0$, where $t(x)$ is a polynomial of degree at most 6 in x . This polynomial must vanish at all of the simultaneous zeros of $q(x, y)$ and $s(x, y)$. We know that there are six such zeros for every μ . However, it is very easy

¹⁷ Klein called it Pascal's theorem.

¹⁸ This polynomial is the difference of two completely factored cubics, by coincidence exactly the kind of polynomial that arises in the six-line locus problem, even though we are not dealing with the distances to any lines here.

to choose μ so that there will be a seventh common zero. With that choice of μ , the polynomial $t(x)$ must have seven zeros, and hence must vanish identically. But since $t(x)$ was the result of eliminating y between the two equations $q(x, y) = 0$ and $s(x, y) = 0$, it now follows that $q(x, y)$ divides $s(x, y)$ (see Problem 12.5 below). That is, the equation $s(x, y) = 0$ can be written as $(ax + by + c)q(x, y) = 0$. Hence its solution set consists of the conic and the line $ax + by + c = 0$, and this line must contain the other three points of intersection.

Conic sections and quadratic functions in general continued to be a source of new ideas for geometers during the early nineteenth century. Plücker liked to use homogeneous coordinates to give a symmetric description of a quadric surface. To take the simplest example, consider the sphere of radius 2 in three-dimensional space with center at $(2, 3, 1)$, whose equation is

$$(x - 2)^2 + (y - 3)^2 + (z - 1)^2 = 4.$$

If x , y , and z , are replaced by ξ/τ , η/τ , and ζ/τ and each term is multiplied by τ^2 , this equation becomes a homogeneous quadratic relation in the four variables (ξ, η, ζ, τ) :

$$(\xi - 2\tau)^2 + (\eta - 3\tau)^2 + (\zeta - \tau)^2 = 4\tau^2.$$

The sphere of unit radius centered at the origin then has the simple equation $\tau^2 - \xi^2 - \eta^2 - \zeta^2 = 0$. Plücker introduced homogeneous coordinates in 1830. One of their advantages is that if $\tau = 0$, but the other three coordinates are not all zero, the point (ξ, η, ζ, τ) can be considered to be located on a sphere of infinite radius. The point $(0, 0, 0, 0)$ is excluded, since it seems to correspond to all points at once.

Homogeneous coordinates correspond very well to the ideas of projective geometry, in which a point in a plane is identified with all the points in three-dimensional space that project to that point from a point outside the plane. If, for example, we take the center of projection as $(0, 0, 0)$ and identify the plane with the plane $z = 1$, that is, each point (x, y) is identified with the point $(x, y, 1)$, the points that project to (x, y) are all points (tx, ty, t) , where $t \neq 0$. Since the equation of a line in the (x, y) -plane has the form $ax + by + c = 0$, one can think of the coordinates (a, b, c) as the coordinates of the line. Here again, multiplication by a nonzero constant does not affect the equation, so that these coordinates can be identified with (ta, tb, tc) for any $t \neq 0$. Notice that the condition for the point (x, y) to lie on the line (a, b, c) is that $\langle (a, b, c), (x, y, 1) \rangle = a \cdot x + b \cdot y + c \cdot 1 = 0$, and this condition is unaffected by multiplication by a constant. The duality between points and lines in a plane is then clear. Any triple of numbers, not all zero, can represent either a point or a line, and the incidence relation between a point and a line is symmetric in the two. We might as well say that the line lies on the point as that the point lies on the line.

Equations can be written in either line coordinates or point coordinates. For example, the equation of an ellipse can be written in homogeneous point coordinates (ξ, η, ζ) as

$$b^2c^2\xi^2 + a^2c^2\eta^2 = a^2b^2\zeta^2,$$

or in line coordinates (λ, μ, ν) as

$$a^2\lambda^2 + b^2\mu^2 = c^2\nu^2,$$

where the geometric meaning of this last expression is that the line (λ, μ, ν) is tangent to the ellipse.

2.11. Arthur Cayley. Homogeneous coordinates provided important invariants and covariants¹⁹ in projective geometry. One such invariant under orthogonal transformations (those that leave the sphere fixed) is the angle between two planes $Ax + By + Cz = D$ and $A'x + B'y + C'z = D'$, given by

$$(2) \quad \arccos \left(\frac{AA' + BB' + CC'}{\sqrt{A^2 + B^2 + C^2} \sqrt{(A')^2 + (B')^2 + (C')^2}} \right).$$

In his "Sixth memoir on quantics," published in the *Transactions of the London Philosophical Society* in 1858, Cayley fixed a "quantic" (quadratic form) $\sum \alpha_{ij} u_i u_j$, whose zero set was a quadric surface that he called the *absolute*, and defined angles by analogy with Eq. (2) and other metric concepts by a similar analogy. In this way he obtained the *general projective metric*, commonly called the *Cayley metric*. It allowed metric geometry to be included in descriptive-projective geometry. As Cayley said, "Metrical geometry is thus a part of descriptive geometry and descriptive geometry is all geometry." By suitable choices of the absolute, one could obtain the geometry of all kinds of quadric curves and surfaces, including the non-Euclidean geometries studied by Gauss, Lobachevskii, Bolyai, and Riemann. Klein (1926, p. 150) remarked that Cayley's models were the most convincing proof that these geometries were consistent.

3. Differential geometry

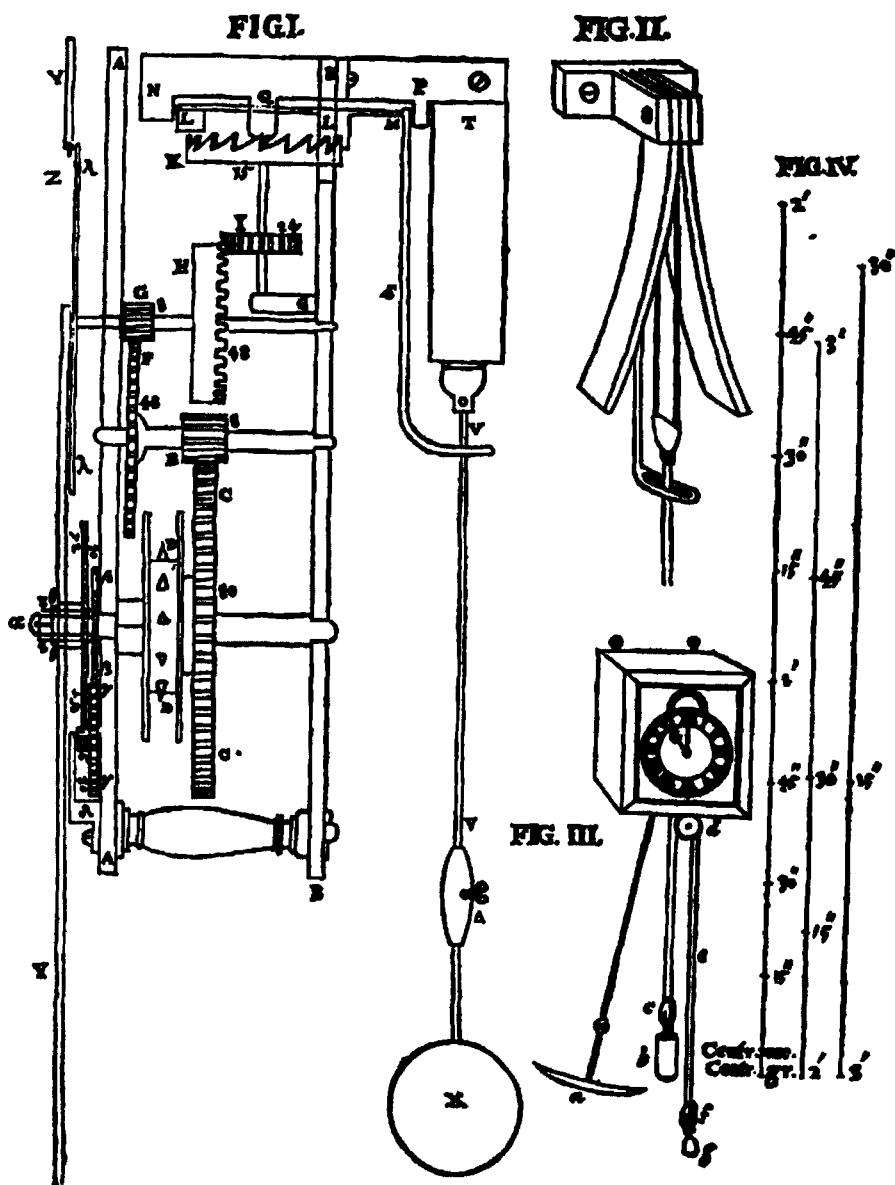
Differential geometry is the study of curves and surfaces (from 1852 onwards, manifolds) using the methods of differential calculus, such as derivatives and local series expansions. This history falls into natural periods defined by the primary subject matter: first, the tangents and curvatures of plane curves; second, the same properties for curves in three-dimensional space; third, the analogous properties for surfaces, geodesics on surfaces, and minimal surfaces; fourth, the application (conformal mapping) of surfaces on one another; fifth, very broad expansions of all these topics, to embrace n -dimensional manifolds and global properties instead of local.

3.1. Huygens. Struik (1933) and Coolidge (1940, p. 319) agree that credit for the first exploration of secondary curves generated by a plane curve—the involute and evolute—occurred in Christiaan Huygens' work *Horologium oscillatorium (Of Oscillating Clocks)* in 1673, even though calculus had not yet been developed. The involute of a curve is the path followed by the endpoint of a taut string being wound onto the curve or unwound from it. Huygens did not give it a name; he simply called it the "line [curve] described by evolution." There are as many involutes as there are points on the curve to begin or end the winding process.

Huygens was seeking a truly synchronous pendulum clock, and he needed a pendulum that would have the same period of oscillation no matter how great the amplitude of the oscillation was.²⁰ Huygens found the mathematically ideal solution of the problem in two properties of the cycloid. First, a frictionless particle

¹⁹ According to Klein (1926, p. 148), the distinction between an invariant and a covariant is not essential. Any algebraic expression that remains unchanged under a family of changes of coordinates is a covariant if it contains variables, and is an invariant if it contains only constants.

²⁰ Despite the legend that Galileo observed a chandelier swinging and noticed that all its swings, whether wide or short, required the same amount of time to complete, for circular arcs that observation is only true approximately for small amplitudes, as anyone who has done the experiment in high-school physics will have learned.



Huygens' cycloidal pendulum, from his *Horologium oscillatorium*.

© Stock Montage, Inc.

requires the same time to slide to the bottom of a cycloid no matter where it begins; second, the involute of a cycloid is another cycloid. He therefore designed a pendulum clock in which the pendulum bob was attached to a flexible leather strap that is confined between two inverted cycloidal arcs. The pendulum is thereby forced to fall along the involute of a cycloid and hence to be truly tautochronous. Reality being more complicated than our dreams, however, this apparatus—like

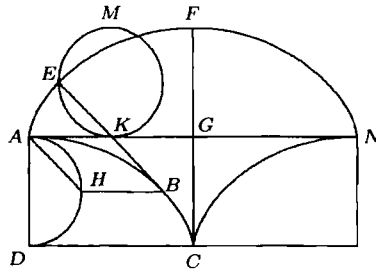


FIGURE 12. Huygens' construction of the "curve formed by evolution" from a cycloid—an identical cycloid.

Dürer's mechanical drawing methods—does not really work any better than the standard methods.²¹

Referring to Fig. 12, in which a line is drawn from one peak of a cycloid to the next and an identical cycloid then drawn atop that line, he showed that BE is perpendicular to the cycloid. But the curve that cuts all the tangents to another curve at right angles is precisely the "curve generated by evolution."

3.2. Newton. In his *Fluxions*, which was first published in 1736, after his death, even though it appears to have been written in 1671, Newton found the circle that best fits a curve. Struik (1933, 19, p. 99) doubted that this material was really in the 1671 manuscript. Be that as it may, the topic occurs as Problem 5 in the *Fluxions*: *At any given Point of a given Curve, to find the Quantity of Curvature.* Newton needed to find a circle tangent to the curve at a given point, which meant finding its center. However, Newton wanted not just any tangent circle. He assumed that if a circle was tangent to a curve at a point and "no other circle can be interscribed in the angles of contact near that point, . . . that circle will be of the same curvature as the curve is of, in that point of contact." In this connection he introduced terms *center of curvature* and *radius of curvature* still used today. His construction is shown in Fig. 13, in which one unnecessary letter has been removed and the figure has been rotated through a right angle to make it fit the page. The weak point of Newton's argument was his claim that, "If CD be conceived to move, while it insists [remains] perpendicularly on the Curve, that point of it C (if you except the motion of approaching to or receding from the Point of Insistence C ,) will be least moved, but will be as it were the Center of Motion." Huygens had had this same problem with clarity. Where Huygens had referred to points that *can be treated as* coincident, Newton used the phrase *will be as it were*.

Newton also treated the problem of the cycloidal pendulum in his *Principia Mathematica*, published in 1687. Huygens had found the evolute of a complete arch of a cycloid. That is, the complete arch is the involute of the portion of two half-arches starting at the halfway point on the arch. In Proposition 50, Problem 33 of Book 1, Newton found the evolute for an arbitrary piece of the arch, which was

²¹ The master's thesis of Robert W. Katsma at California State University at Sacramento in the year 2000 was entitled "An analysis of the failure of Huygens' cycloidal pendulum and the design and testing of a new cycloidal pendulum." Katsma was granted patent 1992-08-18 in Walla Walla County for a cycloidal pendulum. However, the theoretical consensus is that "in every case, such devices would introduce greater errors into the going of a good clock than the errors they are supposed to eliminate." (See the website <http://www.ubr.com/clocks/nawec/hsc/hsn95a.html>.)

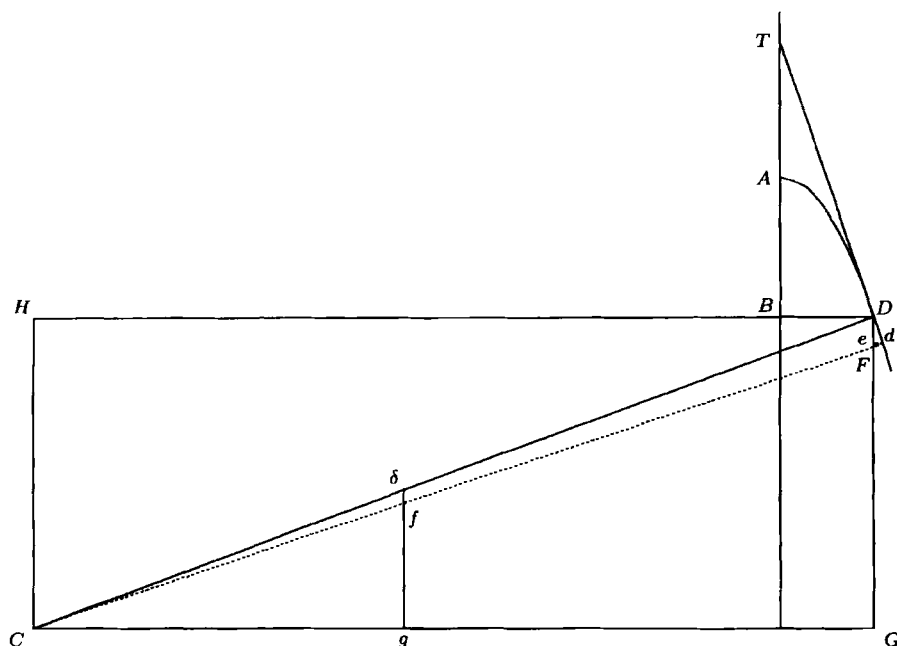


FIGURE 13. Newton's construction of the radius of curvature, from his posthumously published *Fluxions*.

a much more complicated problem. It was, however, once again a cycloid. This evolute made it possible to limit the oscillations of a cycloidal pendulum by putting a complete cycloidal frame in place to stop the pendulum when the thread was completely wound around the evolute.

3.3. Leibniz. Leibniz' contributions to differential geometry began in 1684, when he gave the rules for handling what we now call differentials. His notation is essentially the one we use today. He regarded x and $x + dx$ as infinitely near values of x and v and dv as the corresponding infinitely near values of v on a curve defined by an equation relating x and v . At a maximum or minimum point he noted that $dv = 0$, so that the equation defining the curve had a double root (v and $v + dv$) at that point. He noted that the two cases could be distinguished by the concavity of the curve, defining the curve to be concave if the difference of the increments ddv (which we would now write as $(d^2v/dx^2) dx^2$) was positive, so that the increments dv themselves increased with increasing v . He defined a point where the increments changed from decreasing to increasing to be a *point of opposite turning* (*punctum flexus contrarii*), and remarked that at such a point (if it was a point where $dv = 0$ also), the equation had a triple root. What he said is easily translated into the language of today, by looking at the equation $0 = f(x + h) - f(x)$. Obviously, $h = 0$ is a root. At a maximum or minimum, it is a double root. If the point x yields $dv = 0$ (that is, $f'(x) = 0$) but is not a maximum or minimum, then $h = 0$ is a triple root.

In 1686 he was the first to use the phrase *osculating circle*. He explained the matter thus:

In the infinitely small parts of a curve it is possible to consider not only the direction or inclination or declination, as has been done up to now, but also the change in direction or curvature (*flexura*), and as the measures of the direction of curves are the simplest lines of geometry having the same direction at the same point, that is, the tangent lines, likewise the measure of curvature is the simplest curve having at the same point not only the same direction but also the same curvature, that is a circle not only tangent to the given curve but, what is more, osculating.²²

Leibniz recognized the problem of finding the evolute as that of constructing "not merely an arbitrary tangent to a single curve at an arbitrary point, but a unique common tangent²³ of infinitely many curves belonging to the same order." That meant differentiating with respect to the parameter and eliminating it between the equation of the family and the differentiated equation. In short, Leibniz was the first to discuss what is now called the envelope of a family of curves defined by an equation containing a parameter.

3.4. The eighteenth century. Compared to calculus, differential equations, and analysis in general, differential geometry was not the subject of a large number of papers in the eighteenth century. Nevertheless, there were important advances.

Euler. According to Coolidge (1940, p. 325), Euler's most important contribution to differential geometry came in a 1760 paper on the curvature of surfaces. In that paper he observed that different planes cutting a surface at a point would generally intersect it in curves having different curvatures, but that the two planes for which this curvature was maximal or minimal would be at right angles to each other. For any other plane, making angle α with one of these planes, the radius of curvature would be

$$r = \frac{2fg}{f + g + (g - f) \cos 2\alpha},$$

where f and g are the minimum and maximum radii of curvature at the point. Nowadays, because of an 1813 treatise of Monge's student Pierre Dupin (1784-1873), this formula is written in terms of the curvature $1/r$ as

$$\frac{1}{r} = \frac{\cos^2 \alpha}{g} + \frac{\sin^2 \alpha}{f},$$

where α is the angle between the given cutting plane and the plane in which the curvature is minimal ($1/g$). The equation obviously implies that in a plane perpendicular to the given plane the curvature would be the same expression with the cosine and sine reversed, or, what is the same, with f and g reversed.

Another fundamental innovation due to Euler was the introduction of the now-familiar idea of a parameterized surface, in a 1770 paper on surfaces that can be mapped into a plane. The canvas on which an artist paints and the paper on which an engineer or architect draws plans are not only two-dimensional but also *flat*, having curvature zero. Parameters allow the mathematician or engineer to represent information about any curved surface, as Euler remarked, in the form of

²² Literally, *kissing*.

²³ The tangent was not necessarily to be a straight line.

functions $(t, u) \mapsto (x(t, u), y(t, u), z(t, u))$. Quantities such as curvature and area are then expressed as functions of the parameters (t, u) .

Lagrange. Another study of surfaces, actually a paper in the calculus of variations, was Lagrange's 1762 work on extremal values of integrals.²⁴ The connection with differential geometry is in the problem of minimal surfaces and isoperimetric problems, although he began with the brachystochrone problem (finding the curve of most rapid descent for a falling body). Lagrange found a necessary condition for a surface $z = f(x, y)$ to be minimal.

The French geometers. After these "preliminaries" we finally arrive at the traditional beginning of differential geometry, a 1771 paper of Monge on curves in space and his 1780 paper on curved surfaces. Monge elaborated Leibniz' idea for finding the envelope of a family of lines, considering a family of planes parametrized by their intersections with the z -axis, and obtained the equation of the surface that is the envelope of the family of planes and can be locally mapped into a plane without stretching or shrinking.

3.5. Gauss. With the nineteenth century, differential geometry entered on a period of growth and has continued to reach new heights for two full centuries. The first mathematician to be mentioned is Gauss, who during the 1820s was involved in mapping the region of Hannover in Lower Saxony, where Göttingen is located. This mapping had been ordered by King George IV of England, who was also Elector of Hannover by inheritance from his great grandfather George I. Gauss had been interested in geodesy for many years (Reich, 1977, pp. 29–34) and had written a paper in response to a problem posed by the Danish Academy of Sciences. This paper, which was published in 1825, discussed conformal mapping, that is, mappings that are a pure magnification at each point, so that directions are preserved and the limiting ratio of the actual distance between two points to the map distance between them as one of them approaches the other is the same for approach from any direction.

Involvement with the mapping project inspired Gauss to reflect on the mathematical aspects of developing a curved surface on a flat page and eventually, the more general problem of developing one curved surface on another, that is, mapping the surfaces so that the ratio that the distance from a given point P to a nearby point Q has to the distance between their images P' and Q' tends to 1 as Q tends to P . Gauss apparently planned a full-scale treatise on geodesy but never completed it. Two versions of his major work *Disquisitiones generales circa superficies curvas* (*General Investigations of Curved Surfaces*) were written in the years 1825 and 1827. In the preface to the latter Gauss explained the problem he had set: "to find all representations of a given surface upon another in which the smallest elements remain unchanged." He admitted that some of what he was doing needed to be made more precise through a more careful statement of hypotheses, but wished to show certain results of fundamental importance in the general problem of mapping.

A simple and fruitful technique that Gauss used was to represent any line in space by a point on a fixed sphere of unit radius: the endpoint of the radius parallel to the line.²⁵ This idea, he said, was inspired by the use of the celestial sphere

²⁴ *Œuvres de Lagrange*, T. 1, pp. 335–362.

²⁵ An oriented line is meant here, since there are obviously two opposite radii parallel to the line. Gauss surely knew that the order of the parameters could be used to fix this orientation.

in geometric astronomy. This unit sphere is used in mapping a curved surface by taking the normal line at each point of the surface and mapping it to a point on the sphere, as described, so that the sphere and the surface have parallel normal lines at corresponding points. Obviously a plane maps to a single point under this procedure, since all of its normal lines are parallel to one another. Gauss proposed to use the area of the portion of the sphere covered by this map as a measure of curvature of the surface in question. He called this area the *total curvature* of the surface. He refined this total curvature by specifying that it was to be positive if the surface was convex in both of two mutually perpendicular directions and negative if it was convex in one direction and concave in the other (like a saddle). Gauss gave an informal discussion of this question in terms of the side of the surface on which the normals were to be erected. When the quality of convexity varied in different parts of a surface, Gauss said, a still more refined definition was necessary, which he found it necessary to omit. Along with the total curvature he defined what we would call its density function and he called the *measure of curvature*, namely the ratio of the total curvature of an element of surface to the area of the same element of surface, which he denoted k . The simplest example is provided by a sphere of radius R , any region of which projects to the similar region on the unit sphere. The ratio of the areas is $k = 1/R^2$, which is therefore the measure of curvature of a sphere at every point.

Gauss used two mappings from the parameter space (p, q) into three-dimensional space. The first was the mapping onto the surface itself:

$$(p, q) \mapsto (x(p, q), y(p, q), z(p, q)).$$

The second was the mapping

$$(p, q) \mapsto (X(p, q), Y(p, q), Z(p, q))$$

to the unit sphere, which takes (p, q) to the three direction cosines of the normal to the surface at the point $(x(p, q), y(p, q), z(p, q))$.

From these preliminaries, Gauss was able to derive very simply what he himself described as "almost everything that the illustrious Euler was the first to prove about the curvature of curved surfaces." In particular, he showed that his measure of curvature k was the reciprocal of the product of the two principal radii of curvature that Euler called f and g . He then went on to consider more general parameterized surfaces. Here he introduced the now-standard quantities E , F , and G , given by

$$\begin{aligned} E &= \left(\frac{\partial x}{\partial p}\right)^2 + \left(\frac{\partial y}{\partial p}\right)^2 + \left(\frac{\partial z}{\partial p}\right)^2, \\ F &= \frac{\partial x}{\partial p} \frac{\partial x}{\partial q} + \frac{\partial y}{\partial p} \frac{\partial y}{\partial q} + \frac{\partial z}{\partial p} \frac{\partial z}{\partial q}, \\ G &= \left(\frac{\partial x}{\partial q}\right)^2 + \left(\frac{\partial y}{\partial q}\right)^2 + \left(\frac{\partial z}{\partial q}\right)^2, \end{aligned}$$

and what is now called the first fundamental form for the square of the element of arc length:

$$ds^2 = E dp^2 + 2F dp dq + G dq^2.$$

It is easy to compute that the element of area—the area of an infinitesimal parallelogram whose sides are $(\frac{\partial x}{\partial p} dp, \frac{\partial y}{\partial p} dp, \frac{\partial z}{\partial p} dp)$ and $(\frac{\partial x}{\partial q} dq, \frac{\partial y}{\partial q} dq, \frac{\partial z}{\partial q} dq)$ —is just

$\Delta dp dq$, where $\Delta = \sqrt{EG - F^2}$. Gauss denoted the analogous expression for the mapping $(p, q) \mapsto (X(p, q), Y(p, q), Z(p, q))$, by

$$(3) \quad D dp^2 + 2D' dp dq + D'' dq^2.$$

It turns out that D is just Δ times the cosine of the angle between the normal line to the surface and the line through the origin passing through the point

$$\left(\frac{\partial^2 x}{\partial p^2}, \frac{\partial^2 x}{\partial p \partial q}, \frac{\partial^2 x}{\partial q^2} \right),$$

and similarly for D' and D'' with x replaced by y and z respectively. This coincidence is particular to three-dimensional space, since there just happen to be three different second-order partial derivatives.

The expression in formula (3) is now divided by Δ and the quotient, called the *second fundamental form*, is written $e dp^2 + 2f dp dq + g dq^2$. The element of area on the sphere is $(DD'' - (D')^2) dp dq$. Hence the measure of curvature—what is now called the *Gaussian curvature* and denoted k —is

$$\frac{DD'' - (D')^2}{(EG - F^2)^2},$$

or as it is now written,

$$\frac{eg - f^2}{EG - F^2}.$$

Gauss found another expression for k involving only the quantities E , F , and G and their first and second partial derivatives with respect to the parameters p and q . The expression was complicated, but it was needed for theoretical purposes, not computation.

In a very prescient remark that was later to be developed by Riemann, Gauss noted that “for finding the measure of curvature, there is no need of finite formulæ, which express the coordinates x , y , z as functions of the indeterminates p , q ; but that the general expression for the magnitude of any linear element is sufficient.” The idea is that the geometry of a surface is to be built up from the infinitesimal level using the parameters, not derived from the metric imposed on it by its position in Euclidean space. That is the essential idea of what is now called a *differentiable manifold*.

It is also clear from Gauss' correspondence (Klein, 1926, p. 16) that Gauss already realized that non-Euclidean geometry was consistent. In fact, the question of consistency did not trouble him; he was more interested in measuring large triangles to see if the sum of their angles could be demonstrably less than two right angles. If so, what we now call hyperbolic geometry would be more convenient for physics than Euclidean geometry.

Gauss considered the possibility of developing one surface on another, that is, mapping it in such a way that lengths are preserved on the infinitesimal level. If the mapping is $(x, y, z) \mapsto (u, v, w)$, then by composition, u , v , and w are all functions of the same parameters that determine x , y , and z , and they generate functions E' , F' , and G' for the second surface that must be equal to E , F , and G at the corresponding points, since that is what is meant by developing one surface on another. But since he had just derived an expression for the measure of curvature that depended only on E , F , G and their partial derivatives, he was able to state the profound result that has come to be called his *theorema egregium* (*outstanding theorem*):

If a curved surface is mapped on any other surface, the measure of curvature at each point remains unchanged.

Among other consequences, this meant that surfaces that can be developed on a plane, such as a cone or cylinder, must have Gaussian curvature 0 at each point.

With the first fundamental form Gauss was able to derive a pair of differential equations that must be satisfied by geodesic lines, which he called *shortest lines*,²⁶ and prove that the endpoints of a geodesic circle—the set of geodesics originating at a given point and having a given length—form a curve that intersects each geodesic at a right angle. This result was the foundation for a generalized theory of polar coordinates on a surface, using p as the distance along a geodesic from a variable point to a pole of reference and q as the angle between that geodesic and a fixed geodesic through the pole. This topic very naturally led to the subject of geodesic triangles, formed by joining three points to one another along geodesics. Since he had shown earlier that the element of surface area was

$$d\sigma = \sqrt{EG - F^2} dp dq,$$

and that this expression was particularly simple when one of the sets of coordinate lines consisted of geodesics (as in the case of a sphere, where the lines of longitude are geodesics), the total curvature of such a triangle was easily found for a geodesic triangle and turned out to be

$$A + B + C - \pi,$$

where A , B , and C are the angles of the triangle, expressed in radians. For a plane triangle this expression is zero. For a spherical triangle it is, not surprisingly, the area of the triangle divided by the square of the radius of the sphere. In this way, area, curvature, and the sum of the angles of a triangle were shown to be linked on curved surfaces. This result was the earliest theorem on global differential geometry, since it applies to any surface that can be triangulated. In its modern, developed version, it relates curvature to the topological property of the surface as a whole known as the Euler characteristic. It is called the *Gauss-Bonnet theorem* after Pierre Ossian Bonnet (1819–1892), who introduced the notion of the geodesic curvature of a curve on a surface (that is, the tangential component of the acceleration of a point moving along the curve with unit speed)²⁷ and generalized the formula to include this concept.

3.6. The French and British geometers. In France differential geometry was of interest for a number of reasons connected with physics. In particular, it seemed applicable to the problem of heat conduction, the theory of which had been pioneered by such outstanding mathematicians as Jean-Baptiste Joseph Fourier (1768–1830), Siméon-Denis Poisson (1781–1840), and Gabriel Lamé (1795–1870), since isothermal surfaces and curves in a body were a topic of primary interest. It also applied to the theory of elasticity, studied by Lamé and Sophie Germain, among others. Lamé developed a theory of elastic waves that he hoped would explain light propagation in an elastic medium called ether. Sophie Germain noted that the average

²⁶ According to Klein (1926, Vol. 2, p. 148), the term *geodesic* was first used by Joseph Liouville (1809–1882) in 1850. Klein cites an 1893 history of the term by Paul Stäckel (1862–1919) as source.

²⁷ According to Struik (1933, 20, pp. 163, 165), even this concept was anticipated by Gauss in an unpublished paper of 1825 and followed up on by Ferdinand Minding (1806–1885) in a paper in Crelle's *Journal* in 1830.

of the two principal curvatures derived by Euler would be the same for any two mutually perpendicular planes cutting a surface. She therefore recommended this average curvature as the best measure of curvature. Her approach does indeed make sense in elasticity theory,²⁸ but turns out not to be so useful for pure geometry.²⁹ Joseph Liouville (1809–1882), who founded the *Journal de mathématiques pures et appliquées* in 1836 and edited it until 1874, proved that conformal maps of three-dimensional regions are far less varied than those in two dimensions, being necessarily either inversions or similarities or rigid motions. He published this result in the fifth edition of Monge's book on the applications of analysis to geometry. In contrast, a mapping $(x, y) \mapsto (u, v)$ is conformal if and only if one of the functions $u(x, y) \pm iv(x, y)$ is analytic. As a consequence, there is a rich supply of conformal mappings of the plane.

After Newton differential geometry languished in Britain until the nineteenth century, when William Rowan Hamilton (1805–1865) published papers on systems of rays, building the foundation for the application of differential geometry to differential equations. Another British mathematician, George Salmon (1819–1904), made the entire subject more accessible with his famous textbooks *Higher Plane Curves* (1852) and *Analytic Geometry of Three Dimensions* (1862).

3.7. Riemann. Once the idea of using parameters to describe a surface has been grasped, the development of geometry can proceed algebraically, without reference to what is possible in three-dimensional Euclidean space. This idea was understood by Hermann Grassmann (1809–1877), a secondary-school teacher, who wrote a philosophically inclined mathematical work published in 1844 under the title *Die lineale Ausdehnungslehre, ein neuer Zweig der Mathematik* (*The Theory of Lineal Extensions, a New Branch of Mathematics*). This work, which developed ideas Grassmann had conceived earlier in a work on the ebb and flow of tides, contained much of what is now regarded as multilinear algebra. What we call the coefficients in a linear combination of vectors Grassmann called the numbers by means of which the quantity was derived from the other quantities. He introduced what we now call the tensor product and the wedge product for what he called extensive quantities. He referred to the tensor product simply as the *product* and the wedge product as the *combinatory product*. The tensor product of two extensive quantities $\sum \alpha_r e_r$ and $\sum \beta_s e_s$ was

$$\left[\sum_r \alpha_r e_r, \sum_s \beta_s e_s \right] = \sum_{r,s} \alpha_r \beta_s [e_r, e_s].$$

The combinatory product was obtained by applying to this product the rule that $[e_r, e_s] = -[e_s, e_r]$ (antisymmetrizing). The determinant is a special case of the combinatory product. Grassmann remarked that when the factors are “numerically related” (which we call linearly dependent), the combinatory product would be zero. When the basic units e_r and e_s were entirely distinct, Grassmann called the combinatory product the *outer product* to distinguish it from the *inner product*, which is still called by that name today and amounts to the ordinary dot product

²⁸ In particular, her concept of the average curvature plays a role in the Navier–Stokes equations (<http://www.navier-stokes.net/nsbcst.htm>).

²⁹ However, the average curvature must be zero on a minimal surface.

when applied to vectors in physics. Grassmann remarked that parentheses have no effect on the outer product—in our terms, it is an associative operation.³⁰

Working with these concepts, Grassmann defined the *numerical value* of an extended quantity as the positive square root of its inner square, exactly what we now call the absolute value of a vector in n -dimensional space. He proved that “the quantities of an orthogonal system are not related numerically,” that is, an orthogonal set of nonzero vectors is linearly independent.

Historians of mathematics seem to agree that, because of its philosophical tone and unusual nomenclature, *Ausdehnungslehre* did not attract a great deal of notice until Grassmann revised it and published a more systematic exposition in 1862. If that verdict is correct, there is a small coincidence in Riemann’s use of the term “extended,” which appears to mimic Grassmann’s use of the word, and in his focus on a general number of dimensions in his inaugural lecture at the University of Göttingen. Riemann’s most authoritative biographer Laugwitz (1999, p. 223) says that Grassmann’s work would have been of little use to Riemann, since for him linear algebra was a trivial subject.³¹ This lecture was read in 1854, with the aged Gauss in the audience.³² Although Riemann’s lecture “Über die Hypothesen die der Geometrie zu Grunde liegen” (“On the hypotheses that form the basis of geometry”) occupies only 14 printed pages and contains almost no mathematical symbolism—it was aimed at a largely nonmathematical audience—it set forth ideas that had profound consequences for the future of both mathematics and physics. As Hermann Weyl said:

The same step was taken here that was taken by Faraday and Maxwell in physics, the theory of electricity in particular, ... by passing from the theory of action at a distance to the theory of local action: the principle of understanding the world from its behavior on the infinitesimal level. [Narasimhan, 1990, p. 740]

In the first section Riemann began by developing the concept of an n -fold extended quantity, asking the indulgence of his audience for delving into philosophy, where he had limited experience. He cited only some philosophical work of Gauss and of Johann Friedrich Herbart (1776–1841), a mathematically inclined philosopher whose attempts to quantify sense impressions was an early form of mathematical psychology.³³ He began with the concept of quantity in general, which arises when some general concept can be defined (measured or counted) in different ways. Then, according as there is or is not a continuous transformation from one of the

³⁰ To avoid confusing the reader who knows that the cross product is not an associative product, we note that the outer product applies only when each of the factors is orthogonal to the others. In three dimensional space the cross product of three such vectors, however they are grouped, is always zero.

³¹ One can’t help wondering about the *multilinear algebra* that Grassmann was developing. The recognition of this theory as an essential part of geometry is explicit in Felix Klein’s 1908 work on elementary geometry from a higher viewpoint, but Riemann apparently did not make the connection.

³² At the time of the lecture Gauss had less than a year of life remaining. Yet his mind was still active, and he was very favorably impressed by Riemann’s performance.

³³ Herbart’s 1824 book *Psychologie als Wissenschaft, neu gegründet auf Erfahrung, Metaphysik, und Mathematik* (*Psychology as Science on a New Foundation of Experiment, Metaphysics, and Mathematics*) is full of mathematical formulas involving the strength of sense impressions, manipulated by the rules of algebra and calculus.

ways into another, the various determinations of it form a continuous or discrete manifold. He noted that discrete manifolds (sets of things that can be counted, as we would say) are very common in everyday life, but continuous manifolds are rare, the spatial location of objects of sense and colors being almost the only examples.

The main part of his lecture was the second part, in which he investigated the kinds of metric relations that could exist in a manifold if the length of a curve was to be independent of its position. Assuming that the point was located by a set of n coordinates x_1, \dots, x_n (almost the only mathematical symbols that appear in the paper), he considered the kinds of properties needed to define an infinitesimal element of arc length ds along a curve. The simplest function that met this requirements was

$$ds = \sqrt{\sum a_{ij}(x_1, \dots, x_n) dx_i dx_j},$$

where the coefficients a_{ij} were continuous functions of position and the expression under the square root is always nonnegative. The next simplest case, which he chose not to develop, occurred when the Maclaurin series began with fourth-degree terms. As Riemann said,

The investigation of this more general type, to be sure, would not require any essentially different principles, but it would be rather time-consuming and cast relatively little new light on the theory of space; and moreover the results could not be expressed geometrically.

For the case in which coordinates could be chosen so that $a_{ii} = 1$ and $a_{ij} = 0$ when $i \neq j$, Riemann called the manifold *flat*.

Having listed the kinds of properties space was assumed to have, Riemann asked to what extent these properties could be verified by experiment, especially in the case of continuous manifolds. What he said at this point has become famous. He made a distinction between the infinite and the unbounded, pointing out that while space is always assumed to be unbounded, it might very well not be infinite. Then, as he said, assuming that solid bodies exist independently of their position, it followed that the curvature of space would have to be constant, and all astronomical observation confirmed that it could only be zero. But, if the volume occupied by a body varied as the body moved, no conclusion about the infinitesimal nature of space could be drawn from observations of the metric relations that hold on the finite level. "It is therefore quite conceivable that the metric relations of space are not in agreement with the assumptions of geometry, and one must indeed assume this if phenomena can be explained more simply thereby." Riemann evidently intended to follow up on these ideas, but his mind produced ideas much faster than his frail body would allow him to develop them. He died before his 40th birthday with this project one of many left unfinished. He did, however, send an essay to the Paris Academy in response to a prize question proposed (and later withdrawn): *Determine the thermal state of a body necessary in order for a system of initially isothermal lines to remain isothermal at all times, so that its thermal state can be expressed as a function of time and two other variables.* Riemann's essay was not awarded the prize because its results were not developed with sufficient rigor. It was not published during his lifetime.³⁴

³⁴ Klein (1926, Vol. 2, p. 165) notes that very valuable results were often submitted for prizes at that time, since professors were so poorly paid.

Differential geometry and physics. The work of Grassmann and Riemann was to have a powerful impact on the development of both geometry and physics. One has only to read Einstein's accounts of the development of general relativity to understand the extent to which he was imbued with Riemann's outlook. The idea of geometrizing physics seems an attractive one. The Aristotelian idea of force, which had continued to serve through Newton's time, began to be replaced by subtler ideas developed by the Continental mathematical physicists of the nineteenth century, with the introduction of such principles as conservation of energy and minimal action. In his 1736 treatise on mechanics, Euler had shown that a particle constrained to move along a surface by forces normal to the surface, but on which no forces tangential to the surface act, would move along a shortest curve on the surface. And when he discovered the variational principles that enabled him to solve the isoperimetric problem (see Chapter 17), he applied them to the theory of elasticity and vibrating membranes. As he said,

Since the material of the universe is the most perfect and proceeds from a supremely wise Creator, nothing at all is found in the world that does not illustrate some maximal or minimal principle. For that reason, there is absolutely no doubt that everything in the universe, being the result of an ultimate purpose, is amenable to determination with equal success from these efficient causes using the method of maxima and minima. [Euler, 1744, p. 245]

It is known that Riemann was searching for a connection between light, electricity, magnetism, and gravitation at this time.³⁵ In 1846, Gauss' collaborator Wilhelm Weber (1804–1891) had incorporated the velocity of light in a formula for the force between two moving charged particles. According to Hermann Weyl (Narasimhan, 1990, p. 741), Riemann did not make any connection between that search and the content of his inaugural lecture. Laugwitz (1999, p. 222), however, cites letters from Riemann to his brother which show that he did make precisely that connection. In any case, four years later Riemann sent a paper³⁶ to the Royal Society in Göttingen in which he made the following remarkable statement:

I venture to communicate to the Royal Society a remark that brings the theory of electricity and magnetism into a close connection with the theory of light and heat radiation. I have found that the electrodynamic effects of galvanic currents can be understood by assuming that the effect of one quantity of electricity on others is not instantaneous but propagates to them with a velocity that is constant (equal to that of light within observational error).

3.8. The Italian geometers. The unification of Italy in the mid-nineteenth century was accompanied by a surge of mathematical activity even greater than the sixteenth-century work in algebra (discussed in Chapter 14). Gauss had analyzed a general surface by using two parameters and introducing six functions: the coefficients of the first and second fundamental forms. The question naturally arises

³⁵ His lecture was given nearly a decade before Maxwell discovered his famous equations connecting the speed of light with the propagation of electromagnetic waves.

³⁶ This paper was later withdrawn, but was published after his death (Narasimhan, 1990, pp. 288–293).

whether a surface can be synthesized from any six functions regarded as the coefficients of these forms. Do they determine the surface, up to the usual Euclidean motions of translation, rotation, and reflection that can be used to move any prescribed point to a prescribed position and orientation? Such a theorem does hold for curves, as was established by two French mathematicians, Jean Frenet (1816–1900) and Joseph Serret (1818–1885), who gave a set of equations—the Frenet–Serret³⁷ equations—determining the curvature and torsion of a curve in three-dimensional spaces. A curve can be reconstructed from its curvature and torsion up to translation, rotation, and reflection. A natural related question is: Which sets of six functions, regarded as the components of the two fundamental forms, can be used to construct a surface? After all, one needs generally only three functions of two parameters to determine the surface, so that the six given by Gauss cannot be independent of one another.

In an 1856 paper, Gaspare Mainardi (1800–1879) provided consistency conditions in the form of four differential equations, now known as the Mainardi–Codazzi equations,³⁸ that must be satisfied by the six functions $E, F, G, D, D',$ and D'' if they are to be the components of the first and second fundamental forms introduced by Gauss. Mainardi had learned of Gauss' work through a French translation, which had appeared in 1852. These same equations were discovered by Delfino Codazzi (1824–1875) two years later, using an entirely different approach, and helped him to win a prize from the Paris Academy of Sciences. Codazzi published these equations only in 1883.

When Riemann's lecture was published in 1867, the year after his death, it became the point of departure for a great deal of research in Italy.³⁹ One who worked to develop these ideas was Riemann's friend Enrico Betti (1823–1892), who tried to get Riemann a chair of mathematics in Palermo. These ideas led Betti to the notion of the connectivity of a surface. On the simplest surfaces, such as a sphere, every closed curve is the boundary of a region. On a torus, however, the circles of latitude and longitude are not boundaries. These ideas belong properly to topology, discussed in the next section. In his fundamental work on this subject, Henri Poincaré named the maximum number of independent non-boundary cycles in a surface the *Betti number* of the surface, a concept that is now generalized to n dimensions. The n th Betti number is the rank of the n th homology group.

Another Italian mathematician who extended Riemann's ideas was Eugenio Beltrami (1835–1900), whose 1868 paper on spaces of constant curvature contained a model of a three-dimensional space of constant negative curvature. Beltrami had previously given the model of a pseudosphere, as explained in Chapter 11, to represent the hyperbolic plane. It was not obvious before his work that three-dimensional hyperbolic geometry and a three-dimensional manifold of constant negative curvature were basically the same thing. Beltrami also worked out the appropriate n -dimensional analogue of the Laplacian $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}$, which plays a fundamental role in mathematical physics. By working with an integral considered earlier by

³⁷ Frenet gave six equations for the direction cosines of the tangent and principal normal to the curve and its radius of curvature. Serret gave the full set of nine now called by this name, which are more symmetric but contain no more information than the six of Frenet.

³⁸ The Latvian mathematician Karl Mikhailovich Peterson (1828–1881) published an equivalent set of equations in Moscow in 1853, but they went unnoticed for a full century.

³⁹ Riemann went to Italy for his health and died of tuberculosis in Selasca. He was in close contact with Italian mathematicians and even published a paper in Italian.

Jacobi (see Klein, 1926, Vol. 2, p. 190), Beltrami arrived at the operator

$$\Delta u = \frac{1}{\sqrt{a}} \sum_{i=1}^n \frac{\partial}{\partial x^i} \left(\sqrt{a} \sum_{j=1}^n a_{ij} \frac{\partial u}{\partial x^j} \right),$$

where, with the notation slightly modernized, the Riemannian metric is given by the usual $ds^2 = \sum_{i,j=1}^n a_{ij} dx^i dx^j$, and a denotes the determinant $\det(a_{ij})$. The generalized operator is now referred to as the Laplace–Beltrami operator on a Riemannian manifold.

The algebra of Grassmann and its connection with Riemann's general metric on an n -dimensional manifold was not fully codified until 1901, in “Méthodes de calcul différentiel absolu et leurs applications” (“Methods of absolute differential calculus and their applications”), published in *Mathematische Annalen* in 1901, written by Gregorio Ricci-Curbastro (1853–1925) and Tullio Levi-Civita (1873–1941). This article contained the critical ideas of tensor analysis as it is now taught. The absoluteness of the calculus consisted in the great generality of the transformations that it permitted, showing how differential forms changed when coordinates were changed. Although Ricci-Curbastro competed in a prize contest sponsored that year by the Accademia dei Lincei, he was not successful, as some of the judges regarded his absolute differential calculus as “useful but not essential”⁴⁰ to the development of mathematics—the same sort of criticism leveled by Weierstrass against the work of Hamilton in quaternions (see Section 2 of Chapter 15).

The following year Luigi Bianchi (1873–1928) published “Sui simboli a quattro indici e sulla curvatura di Riemann” (“On the quadruply-indexed symbols and Riemannian curvature”), in which he gave the relations among the covariant derivatives of the Riemann curvature tensor, which, however, he derived by a direct method for manifolds of constant curvature, not following the route of Ricci-Curbastro and Levi-Civita. The Bianchi identity was later to play a crucial role in general relativity, assuring local conservation of energy when Einstein's gravitational equation is assumed.

4. Topology

Projections distort the shape of geometric objects, so that some metric properties are lost. Some properties, such as parallelism, however, remain simply because the number of intersections of two curves does not change. The study of space focusing on such very general properties as connections and intersections has been known by various names over the centuries. Latin has two words, *locus* and *situs*, meaning roughly *place* and *position*. The word *locus* is one that we still use today to denote the path followed by a point moving subject to stated constraints. It was the translation of the Greek word *tópos* used by Pappus for the same concept. Since *locus* was already in use, Leibniz fastened on *situs* and mentioned the need for a geometry or analysis of *situs* in a 1679 letter to Huygens.⁴¹ The meaning of *geometria situs* and *analysis situs* evolved gradually. It seems to have been Johann Benedict Listing (1808–1882) who, some time during the 1830s, realized that the Greek root

⁴⁰ See the article on Ricci-Curbastro's paper at <http://www.math.unifi.it/matematicaitaliana/>.

⁴¹ This letter was published in Huygens' *Œuvres complètes*, M. Nijhoff, La Haye, 1888, Vol. 8, p. 216. From the context it appears that Leibniz was calling for some simple way of expressing position “as algebra expresses magnitude.” If so, perhaps we now have what he wanted in the form of vector analysis.

was available. The word *topology* first appeared in the title of his 1848 book *Vorstudien zur Topologie (Prolegomena to Topology)*. Like geometry itself, topology has bifurcated several times, so that one can now distinguish combinatorial, algebraic, differential, and point-set topology.

4.1. Early combinatorial topology. The earliest result that deals with the combinatorial properties of figures is now known as the *Euler characteristic*, although Descartes is entitled to some of the credit.⁴² In a work on polyhedra that he never published, Descartes defined the solid angle at a vertex of a closed polyhedron to be the difference between 2π and the sum of the angles at that vertex. He asserted that the sum of the solid angles in any closed polyhedron was exactly eight right angles. (In our terms, that number is 4π , the area of a sphere of unit radius.) Descartes' work was found among his effects after he died. By chance Leibniz saw it a few decades later and made a copy of it. When it was found among Leibniz' papers, it was finally published. In the eighteenth century, Euler discovered this same theorem in the form that the sum of the angles at the vertices of a closed polyhedron was $4n - 8$ right angles, where n is the number of vertices. Euler noted the equivalent fact that the number of faces and vertices exceeded the number of edges by 2. That is the formula now generally called *Euler's formula*:

$$V - E + F = 2.$$

Somewhat peripheral to the general subject of topology was Euler's analysis of the famous problem of the seven bridges of Königsberg in 1736. In Euler's day there were two islands in the middle of the River Pregel, which flows through Königsberg (now Kaliningrad, Russia). These islands were connected to each other by a bridge, and one of them was connected by two bridges to each shore, the other by one bridge to each shore. The problem was to go for a walk and cross each bridge exactly once, returning, if possible to the starting point. In fact, as one can easily see, it is impossible even to cross each bridge exactly once without boating or swimming across the river. Returning to the starting point merely adds another condition to a condition that is already impossible to fulfill. Euler proved this fact by labeling the two shores and the two islands A , B , C , and D , and representing a stroll as a "word," such as $ABCBD$, in which the bridges are "between" the letters. He showed that any such path as required would have to be represented by an 8-letter word containing three of the letters twice and the other letter three times, which is obviously impossible. This topic belongs to what is now called graph theory; it is an example of the problem of unicursal tracing.

4.2. Riemann. The study of analytic functions of a complex variable turned out to require some concepts from topology. These issues were touched on in Riemann's 1851 doctoral dissertation at Göttingen, "Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse" ("Foundations for a general theory of functions of a complex variable"). Although all analytic functions of a complex variable, both algebraic and transcendental, were encompassed in Riemann's ideas, he was particularly interested in algebraic functions, that is functions $w = f(z)$ that satisfy a nontrivial polynomial equation $p(z, w) = 0$. Algebraic functions are essentially and unavoidably multivalued. To take the simplest example,

⁴² Much of the information in this paragraph is based on the following website: <http://www.math.sunysb.edu/~tony/whatsnew/column/descartes-0899/descartes2.html>.

in which $z - w^2 = 0$, every complex number $z = a + bi$ has two distinct complex square roots:

$$w = \pm(u + iv), \text{ where } u = \sqrt{\frac{\sqrt{a^2 + b^2} + a}{2}} \text{ and } v = \operatorname{sgn}(b) \sqrt{\frac{\sqrt{a^2 + b^2} - a}{2}}.$$

The square roots of the positive real numbers that occur here are assumed positive. There is no way of choosing just one of the two values at each point that will result in a continuous function $w = \sqrt{z}$. In particular, it is easy to show that any such choice must have a discontinuity at some point of the circle $|z| = 1$.

One way to handle this multivaluedness was to take two copies of the z -plane, labeled with subscripts as z_1 and z_2 and place one of the square roots in one plane and the other in the other. This technique was used by Cauchy and had been developed into a useful way of looking at complex functions by Victor Puiseux (1820–1883) in 1850. Indeed, Puiseux seems to have had the essential insights that can be found in Riemann's work, although differently expressed. Riemann is known to have seen the work of Puiseux, although he did not cite it in his own work. He generally preferred to work out his own way of doing things and tended to ignore earlier work by other people. In any case, the essential problem with choosing one square root and sticking to it is that a single choice cannot be continuous on a closed path that encloses the origin without going through it. At some point on such a path, there will be nearby points at which the function assumes two values that are close to being negatives of each other.

Riemann had the idea of cutting the two copies of the z -plane along a line running from zero to infinity (both being places where there is only one square root, assuming a bit about complex infinity). Then if the lower edge of each plane is imagined as being glued to the upper edge of the other,⁴³ the result is a single connected surface in which the origin belongs to both planes. On this new surface a continuous square-root function can be defined. It was the gluing that was really new here. Cauchy and Puiseux both had the idea of cutting the plane to keep a path from winding around a branch point and of using different copies of the plane to map different branches of the function.

Riemann introduced the idea of a *simply connected surface*, one that is disconnected by any cut from one boundary point to another that passes through its interior without intersecting itself. He stated as a theorem that the result of such a cut would be two simply connected surfaces. In general, when a connected surface is cut by a succession of such *crosscuts*, as he called them, the difference between the number of crosscuts and the number of connected components that they produce is a constant, called the *order of connectivity* of the surface. A sphere, for example, can be thought of as a square with adjacent edges glued together, as in Fig. 14. It is simply connected because a diagonal cut disconnects it. The torus, on the other hand, can be thought of as a square with opposite edges identified (see Fig. 14). To disconnect this surface, it is necessary to cut it at least twice, for example, either by drawing both diagonals or by cutting it through its midpoint with two lines parallel to the sides. No single cut will do. The torus is thus doubly connected.

⁴³ You can visualize this operation being performed if you imagine one copy of the plane picked up and turned upside down above the other so that the upper edge is glued to the upper edge and the lower to the lower.

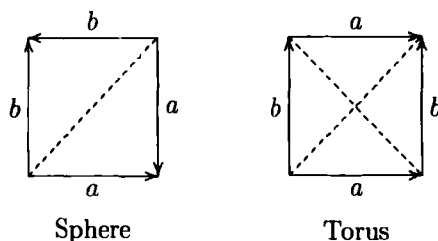


FIGURE 14. Left: The sphere, regarded as a square with edges identified, is disconnected by a diagonal. Right: The torus requires two cuts to disconnect.

4.3. Möbius. One fact that had been thought well established about polyhedra was that in any polyhedron it was possible to direct the edges in such a way that one could trace around the boundary of each face by following the prescribed direction of its edges. Each face would be always to the left or always to the right as one followed the edges around it while looking at it from outside the polyhedron. This fact was referred to as the *edge law* (*Gesetz der Kanten*). The first discovery of a closed polyhedron that violated this condition⁴⁴ was due to Möbius, sometime during the late 1850s. Möbius did not publish this work, although he did submit some of it to the Paris Academy as his entry to a prize competition in 1858. This work was edited and introduced by Curt Reinhardt (dates unknown) and published in Vol. 2 of Möbius' collected works. There in the first section, under the heading "one-sided polyhedra," is a description of the Möbius band as we now know it (Fig. 14). After describing it, Möbius went on to say that although a triangulated polyhedron whose surface was two-sided will apparently contain only two-sided bands, *nevertheless a triangulated polyhedron with a one-sided surface can contain both one- and two-sided bands*.

Möbius explored polyhedra and made a classification of them according to the number of boundary curves they possessed. He showed how more complicated polyhedra could be produced by gluing together a certain set of basic figures. He found an example of a triangulated polyhedron consisting of 10 triangles, six vertices, and 15 edges, rather than 14, as would be expected from Euler's formula for a closed polyhedron: $V - E + F = 2$. This figure is the projective plane, and cannot be embedded in three-dimensional space. If one of the triangles is removed, the resulting figure is the Möbius band, which can be embedded in three-dimensional space.

4.4. Poincaré's *Analysis situs*. Poincaré seemed to be dealing constantly with topological considerations in his work in both complex function theory and differential equations. To set everything that he discovered down in good order, he wrote a treatise on topology called *Analysis situs* in 1895, published in the *Journal de l'École Polytechnique*, that has been regarded as the founding document of modern algebraic topology.⁴⁵ He introduced the notion of homologous curves—curves that (taken together) form the boundary of a surface. This notion could be formalized, so that one could consider formal linear combinations (now called

⁴⁴ In fact, a closed nonorientable polyhedron cannot be embedded in three-dimensional space, so that the edge law is actually true for *closed* polyhedra in three-dimensional space.

⁴⁵ Poincaré followed this paper with a number of supplements over the next decade.

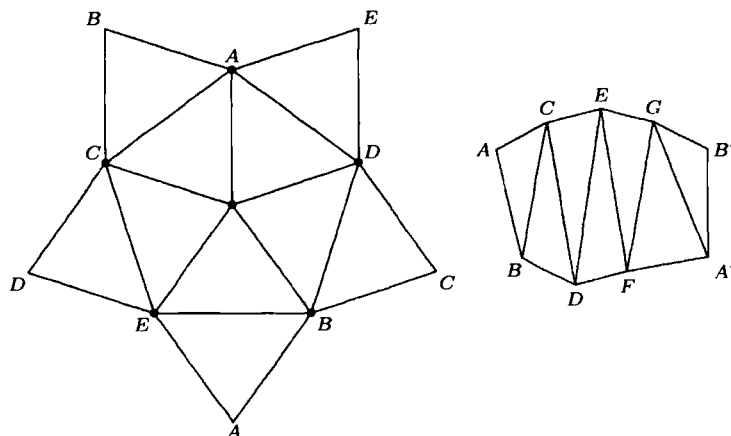


FIGURE 15. Left: the projective plane triangulated and cut open. If two opposite edges with corresponding endpoints are glued together, the figure becomes a Möbius band. In three-dimensional space it is not possible to glue all the edges together as indicated. Right: the Möbius band as originally described by Möbius.

chains) $C = n_1C_1 + \cdots + n_rC_r$ of oriented curves C_i with integer coefficients n_i . The interpretation of such a combination came from analysis: A line integral over C was interpreted as the number $I = n_1I_1 + \cdots + n_rI_r$, where I_j was the line integral over C_j . When generalized to k -dimensional manifolds (called *varieties* by Poincaré) and combined with the concept of the boundary of an oriented manifold as a cycle, this idea was the foundation of homology theory: The k -cycles (k -chains whose boundaries are the zero $(k-1)$ -chain—Poincaré called them *closed varieties*) form a group, of which the k -cycles that are the boundary of a $(k+1)$ -cycle form a subgroup. When two homologous cycles (cycles whose difference is a boundary) are identified, the resulting classes of cycles form the k th *homology group*. For example, in the sphere shown in Fig. 14, the diagonal that is drawn forms a cycle. This cycle is the complete boundary of the upper and lower triangles in the figure, and it turns out that any cycle on the sphere is a boundary. The first homology group of the sphere is therefore trivial (consists of only one element). For the torus depicted in Fig. 14, a and b are each cycles, but neither is a boundary, nor is any cycle $ma + nb$. On the other hand, the cycle formed by adding either diagonal to $a + b$ is the boundary of the two triangles with these edges. Thus, the first homology group of the torus can be identified with the set of cycles $ma + nb$. Any other cycle will be homologous to one of these.

Poincaré also introduced the notion of the *fundamental group* of a manifold. He had been led to algebraic topology partly by his work in differential equations. He discovered the fundamental group by imagining functions satisfying a set of differential equations and being permuted as a point moved around a closed loop. He was thus led to consider formal sums of loops starting and ending at a given point, two loops being equivalent if tracing them successively left the functions invariant. The resulting set of permutations was what he called the *fundamental group*. He cautioned that, despite appearances, the fundamental group was not the same thing as the first homology group, since there was no base point involved in the

homology group. Moreover, he noted, while the order in which the cycles in a chain were traversed was irrelevant, the fundamental group was not necessarily commutative. He suggested redefining the term *simply connected* to mean having a trivial fundamental group. He gave examples to show that the homology groups do not determine the topological nature of a manifold, exhibiting three three-dimensional manifolds all having the same homology groups, but different fundamental groups and therefore not topologically the same (homeomorphic). He then asked a number of questions about fundamental groups, one of which has become famous. *Given two manifolds of the same number of dimensions having the same fundamental group, are they homeomorphic?* Like Fermat's last theorem, this question has been attacked by many talented mathematicians, and proofs have been proposed for a positive answer to the question, but—at least until recently—all such proofs have been found wanting.⁴⁶

4.5. Point-set topology. Topology is sometimes popularly defined as “rubber-sheet geometry,” in the sense that the concepts it introduces are invariant under moving and stretching, provided that no tearing takes place. In the kinds of combinatorial topology just discussed, those concepts usually involve numbers in some form or other—the number of independent cycles on a manifold, the Euler characteristic, and so forth. But there are also topological concepts not directly related to number.

Continuity and connectedness. The most important of these is the notion of *connectedness* or *continuity*. This word denotes a deep intuitive idea that was the source of many paradoxes in ancient times, such as the paradoxes of Zeno. As we shall see in Chapter 15, it is impossible to prove the fundamental theorem of algebra without this concept.⁴⁷ For analysts, it was crucial to know that if a certain function was negative at one point on a line and positive at another, it must assume the value zero at some point between the two points. That property eventually supplanted earlier definitions of continuity, and the property now taken as the definition of continuity is designed to make this proposition true. The clarification of the ideas surrounding continuity occurred in the early part of the nineteenth century and is discussed in more detail in Chapter 17. Once serious analysis of this concept was undertaken, it became clear that many intuitive assumptions about the connectedness of curves and surfaces had been made from the beginning of deductive geometry. These continuity considerations complicated the theory of functions of a real variable for some decades until adequate explanations of it were found. A good example of such problems is provided by Dedekind's construction of the real numbers, discussed in Chapter 8, which he presented as a solution to the problem of defining what is meant by a continuum.

⁴⁶ As of this writing, evidence begins to accumulate that the Russian mathematician Grigori Perlman of the Steklov Institute in St. Petersburg has settled the Poincaré conjecture (Associated Press, January 7, 2004). As a graduate student at Princeton in 1964, when a mathematician came to town claiming to have proved this elusive result, I discussed it with Norman Steenrod (1910–1971), one of the twentieth century's greatest topologists. He told me that proving the conjecture, although difficult, would be a rather uninteresting thing to do, since it would only confirm what people already thought was true. It would have been much more exciting to disprove it.

⁴⁷ Even the second of the four proofs that Gauss gave, which is generally regarded as a purely algebraic proof, required the assumption that an equation of odd degree with real coefficients has a real solution—a fact that relies on connectedness.

Compactness. Another basic concept of point-set topology is that of *compactness*. This concept is needed to make the distinction between being bounded and having a minimum or maximum. The concepts of compactness, connectedness, and continuity are used together nowadays to prove such theorems as Rolle's theorem in calculus.

At least three lines of thought led to the notion of compactness. The first was the search for maxima and minima of functions, that is, points at which the function assumed the largest or smallest possible value. It was clear that a sequence of points x_n could always be found such that $f(x_n)$ tended to a maximum value; that was what a maximum value meant. But did the sequence x_n itself, or some subsequence of it, also converge to a point x ? If so, it was clear from the definition of continuity that x must be a maximum or minimum. This property was studied by the Czech mathematician Bernard Bolzano (1781–1848), who was looking for a proof of the continuity property discussed above. He showed as early as 1817 (see Manheim, 1964, p. 67) that the continuity property could be made to follow from the property that a set of numbers that is bounded above has a least upper bound. He phrased this statement differently, of course, saying that if there is a property possessed by a function at some points, but not all, and that property holds for all points less than some a , there is a smallest number U such that the property holds for all numbers less than U . Bolzano proved this fact by repeated bisection of an interval such that the property holds at the lower endpoint but not the upper. Some 50 years later, after defining real numbers as sequences of rational numbers (with a suitable notion of equivalent sequences), Weierstrass used arguments of this type to deduce that a bounded sequence of real numbers has a convergent subsequence. This theorem, in several closely equivalent forms, is now known as the *Bolzano–Weierstrass theorem*.

The second line of thought leading to compactness was the now-familiar distinction between pointwise continuity and uniform continuity. This distinction was brought to the fore in the mid-1850s, and Dirichlet proved that on an interval $[a, b]$ (including the endpoints) a continuous function was uniformly continuous. He was really the first person to use the idea of replacing a covering by open sets with a finite subcovering. The same theorem was proved by Eduard Heine (1821–1881) in 1872; as a result, Heine found his name attached to one form of the basic theorem.

The third line was certain work in complex analysis by Émile Borel (1871–1956) in the 1890s. Borel was studying analytic continuation, whereby a complex-valued function is expanded as a power series about some point:

$$f(z) = f(z_0) + \sum_{n=1}^{\infty} \frac{f^{(n)}(z_0)}{n!} (z - z_0)^n.$$

If the series has a finite radius of convergence, it represents $f(z)$ only inside a disk. However, it enables all the derivatives of $f(z)$ to be computed at all points of the disk, so that if one forms the analogous series at some point z_1 in the disk different from z_0 , it is possible that the new series will converge at some points outside the original disk. In this way, one can continue a function uniquely along a path γ from one point a to another point b , provided there is such a series with a positive radius of convergence at each point of γ . What is needed is some way of proving that only a finite number of such disks will be required to cover the whole curve γ . The resulting covering theorem was further refined by a number of mathematicians,

including Henri Lebesgue (1875–1941), and is now generally known as the *Heine–Borel theorem*. The word *compact* was first used in 1906 in two equivalent senses in two different papers, by Maurice Fréchet (1878–1973).

Closed and open sets. The word *set*, without which modern mathematicians would not be able to talk at all, was not introduced formally until the 1870s. The history of set theory is discussed in more detail in Chapter 19. At present we merely mention that the idea of a closed set arose from consideration of the set of limit points of a given set (its derived set). In an 1884 paper, Georg Cantor (1854–1918) called a set *closed* if it contained all of its limit points. Since it was easy to show that a limit point of limit points of a set P is itself a limit point of P , it followed that the derived set P' is always a closed set.

Although the phrase *closed set* appears in 1884, its dual—the phrase *open set*—did not appear for nearly two more decades. Weierstrass had used the *concept* of an open set in discussing analytic functions, since he used power series, which required that the function be defined a small disk called a *neighborhood* about each point in its domain of definition. Weierstrass used the German term *Gebiet* (*region*) for such a domain of definition. The phrase *open set* seems to have been used for the first time by W. H. Young in 1902.⁴⁸ In a 1905 paper in descriptive function theory (that is, discussing what it means for a function to be “analytic” in a very general sense), Henri Lebesgue referred specifically to *ensembles ouverts* (*open sets*) and defined them to be the complements of closed sets.

Metric spaces. The frequent repetition of certain basic patterns of reasoning, and perhaps just a normal human penchant for order, led to the creation of very abstract structures around the beginning of the twentieth century. The kind of continuity argument we now associate with δ 's and ε 's was generalized in 1905 by Maurice Fréchet, who considered abstract sets on which there was a sort of distance between two points A and B , denoted (A, B) . This distance had the properties normally associated with distance, that is, $(A, B) = (B, A)$ (the distance from A to B is the same as the distance from B to A), $(A, B) > 0$ if $A \neq B$, and $(A, A) = 0$. Further, he assumed that there was a real-valued function $f(t)$ tending to 0 as t tends to 0, and such that $(A, C) < f(\varepsilon)$ if $(A, B) < \varepsilon$ and $(B, C) < \varepsilon$. Such a structure is now called a *metric space*, although the definition is streamlined somewhat, the third property being replaced by the *triangle inequality*. It can be shown that for each distance function introduced by Fréchet there is an equivalent metric in the modern sense.

In 1906 Fréchet also gave two definitions of the term *compact* (for metric spaces) in the modern sense. In one paper he defined a space to be compact if every infinite subset of it had at least one limit point. In the other he defined compactness to mean that every decreasing sequence of nonempty closed sets had a nonempty intersection. Thus he used both the Bolzano–Weierstrass property and the Heine–Borel property (which are equivalent for metric spaces).

General topology. The notion of a topological space in the modern sense arose in 1914 in the work of the Youngs and in the work of Felix Hausdorff (1868–1942), who was at the time a professor at Bonn. Hausdorff's influential book *Grundzüge der Mengenlehre* (*Elements of Set Theory*) was translated into many languages.

⁴⁸ The early papers of W. H. Young and his wife G. C. Young were published under his name alone, as mentioned in Chapter 4.

The first part of the book is an exposition of abstract set theory as it existed at the time, including cardinal and ordinal numbers, and the early stages of what is now called *descriptive set theory*, that is, the classification of sets according to their complexity, starting with a ground class consisting of closed sets and open sets, and then proceeding up a hierarchy by passing to countable unions and intersections. He invented the term *ring* for a class of sets that was closed under finite unions and intersections and *field* for a class that was closed under set differences and finite unions, but warned in a footnote that “the expressions *ring* and *field* are taken from the theory of algebraic numbers based on an approximate analogy that it will not do to push too far.”⁴⁹

Hausdorff introduced metric spaces, being the first to use that name for them, via the axioms now used, then gave a set of “neighborhood axioms” for a more general type of space:

1. To each point x there corresponds at least one neighborhood U_x ; every neighborhood U_x contains the point x .
2. If U_x and V_x are two neighborhoods of the point x , there is another neighborhood W_x of x contained in both of them.
3. If the point y lies in U_x , there is a neighborhood U_y contained in U_x .
4. For any distinct points x and y , there are two neighborhoods U_x and U_y whose intersection is empty.

These were Hausdorff’s axioms for topology, and they were well designed for discussing the local behavior of functions on a highly abstract level. A quarter-century later, the group of French authors known collectively as Nicolas Bourbaki introduced a global point of view, defining a topological space axiomatically as we know it today, in terms of open sets. The open sets of a space can be any collection that has the empty set and the whole set as members and is closed under arbitrary unions and finite intersections. In those terms, one of Hausdorff’s neighborhoods U_x is any open set O with $x \in O$. Conversely, given a set on which the first three of Hausdorff’s axioms hold, it is easy to show that the sets that are neighborhoods of all of their points form a topology in the sense of Bourbaki. Bourbaki omitted the last property specified by Hausdorff. Spaces having this extra property are now called (appropriately enough) Hausdorff spaces.

Questions and problems

- 12.1.** Judging from Descartes’ remarks on mechanically drawn curves, should he have admitted the conchoid of Nicomedes among the legitimate curves of geometry?
- 12.2.** Prove Menelaus’ theorem and its converse. What happens if the points E and F are such that $AD : AE :: BD : BE$? (Euclid gave the answer to this question.)
- 12.3.** Use Menelaus’ theorem to prove that two medians of a triangle intersect in a point that divides each in the ratio of 1:2.

⁴⁹ The word *ring* in the abstract algebraic sense was also introduced in 1914, in a paper of A. Fraenkel (see Section 2 of Chapter 15). A very influential work in measure theory, written in 1950 by Paul Halmos (b. 1916), caused Hausdorff’s *ring* to fall into disuse and appropriated the term *ring* to mean what Hausdorff called a *field*. Halmos reserved the term *algebra* for a ring, one of whose elements was the entire space. Probabilists, however, use the term *field* for what Halmos called an algebra.

12.4. Deduce Brianchon's theorem for a general conic from the special case of a circle. How do you interpret the case of a regular hexagon inscribed in a circle?

12.5. Fill in the details of Plücker's proof of Brianchon's theorem, as follows: Suppose that the equation of the conic is $q(x, y) = y^2 + r_1(x)y + r_2(x) = 0$, where $r_1(x)$ is a linear polynomial and $r_2(x)$ is quadratic. Choose coordinate axes not parallel to any of the sides of the inscribed hexagon and such that the x -coordinates of all of its vertices will be different, and also choose the seventh point to have x -coordinate different from those of the six vertices. Then suppose that the polynomial generated by the three lines is $s(x, y) = y^3 + t_1(x)y^2 + t_2(x)y + t_3(x) = 0$, where $t_j(x)$ is of degree j , $j = 1, 2, 3$. Then there are polynomials $u_j(x)$ of degree j , $j = 1, 2, 3$, such that

$$s(x, y) = q(x, y)(y - u_1(x)) + (u_2(x)y + u_3(x)).$$

We need to show that $u_2 \equiv 0$ and $u_3 \equiv 0$. At the seven points on the conic where both $q(x, y)$ and $s(x, y)$ vanish it must also be true that $u_2(x)y + u_3(x) = 0$. Rewrite the equation $q(x, y) = 0$ at these seven points as

$$(u_2y)^2 + r_1u_2(u_2y) + u_2^2r_2 = 0$$

observe that at these seven points $u_2y = -u_3$, so that the polynomial $u_3^2 - r_1u_2u_3 + u_2^2r_2$, which is of degree 6, has seven distinct zeros. It must therefore vanish identically, and that means that

$$(2u_3 - r_1u_2)^2 = u_2^2(r_1^2 - 4r_2).$$

This means that either u_2 is identically zero, which implies that u_3 also vanishes identically, or else u_2 divides u_3 . Prove that in the second case the conic must be a pair of lines, and give a separate argument in that case.

12.6. Consider the two equations

$$\begin{aligned} xy &= 0, \\ x(y-1) &= 0. \end{aligned}$$

Show that these two equations are independent, yet have infinitely many common solutions. What kind of conic sections do these equations represent?

12.7. Consider the general cubic equation

$$Ax^3 + Bx^2y + Cxy^2 + Dy^3 + Ex^2 + Fxy + Gy^2 + Hx + Iy + J = 0,$$

which has 10 coefficients. Show that if this equation is to hold for the 10 points $(1, 0), (2, 0), (3, 0), (4, 0), (0, 1), (0, 2), (0, 3), (1, 1), (2, 2), (1, -1)$, all 10 coefficients A, \dots, J must be zero. In general, then, it is not possible to pass a curve of degree 3 through any 10 points in the plane. Use linear algebra to show that it is always possible to pass a curve of degree 3 through any nine points, and that the curve is generally unique.

On the other hand, two *different* curves of degree 3 generally intersect in 9 points, a result known as Bézout's theorem after Étienne Bézout (1730–1783), who stated it around 1758, although Maclaurin had stated it earlier. How does it happen that while nine points generally determine a *unique* cubic curve, yet *two distinct* cubic curves generally intersect in nine points? [Hint: Suppose that a set of eight points $\{(x_j, y_j) : j = 1, \dots, 8\}$ is given for which the system of equations for A, \dots, J has rank 8. Although the system of linear equations for the coefficients is generally of rank 9 if another point is adjoined to this set, there generally is a point

(x_9, y_9) , the ninth point of intersection of two cubic curves through the other eight points, for which the rank will remain at 8.]

12.8. Find the Gaussian curvature of the hyperbolic paraboloid $z = (x^2 - y^2)/a$ at each point using x and y as parameters.

12.9. Find the Gaussian curvature of the pseudosphere obtained by revolving a tractrix about the x -axis. Its parameterization can be taken as

$$\mathbf{r}(u, v) = \left(u - a \tanh\left(\frac{u}{a}\right), a \operatorname{sech}\left(\frac{u}{a}\right) \cos(v), a \operatorname{sech}\left(\frac{u}{a}\right) \sin(v) \right).$$

Observe that the elements of area on both the pseudosphere and its map to the sphere vanish when $u = 0$. (In terms of the first and second fundamental forms, $E = 0 = g$ when $u = 0$.) Hence curvature is undefined along the circle that is the image of that portion of the parameter space. Explain why the pseudosphere can be thought of as “a sphere of imaginary radius.” Notice that it has a cusp along the circle in which it intersects the plane $x = 0$.

12.10. Prove that the Euler relation $V - E + F = 2$ for a closed polyhedron is equivalent to the statement that the sum of the angles at all the vertices is $(2V - 4)\pi$, where V is the number of vertices. [Hint: Assume that the polyhedron has F faces, and that the numbers of edges on the faces are e_1, \dots, e_F . Then the number of edges in the polyhedron is $E = (e_1 + \dots + e_F)/2$, since each edge belongs to two faces. Observe that a point traversing a polygon changes direction by an amount equal to the exterior angle at each vertex. Since the point returns to its starting point after making a complete circuit, the sum of the *exterior* angles of a polygon is 2π . Since the interior angles are the supplements of the exterior angles, we see that their sum is $e_i\pi - 2\pi = (e_i - 2)\pi$. The sum of all the interior angles of the polyhedron is therefore $(2E - 2F)\pi$.]

12.11. Give an informal proof of the Euler relation $V - E + F = 2$ for closed polyhedra, assuming that every vertex is joined by a sequence of edges to every other vertex. [Hint: Imagine the polyhedron inflated to become a sphere. That stretching will not change V , E , or F . Start drawing the edges on a sphere with a single vertex, so that $V = 1 = F$ and $E = 0$. Show that adding a new vertex by distinguishing an interior point of an edge as a new vertex, or by distinguishing an interior point of a face as a new vertex and joining it to an existing vertex, increases both V and E by 1 and leaves F unchanged, while drawing a diagonal of a face increases E and F by 1 and leaves V unchanged. Show that the entire polyhedron can be constructed by a sequence of such operations.]

Part 5

Algebra

Occasionally, a practical problem arises in which it is necessary to invert a sequence of arithmetic operations. That is, we know the result of the operations but not the data. The best examples of this kind of problem come from geometry, and a typical specimen can be seen in the *sangaku* plaque shown in Color Plate 2. This type of problem is the seed of the area we call algebra, whose development can be conveniently divided into three stages. In the first stage, knowing the procedure followed and the result, one is forced to think in the terms that Pappus referred to as analysis, that is, deducing consequences of the formula until one arrives at the data. The main tool in this analysis is the equation, but equations occur explicitly only after a stock of examples has been accumulated. At the second stage, equations are identified as an object of independent interest, and techniques for solving them are developed. In the third stage, a higher-level analysis of the algorithms for solution leads to the subject we now know as modern algebra. We shall devote one chapter to each of these stages.

CHAPTER 13

Problems Leading to Algebra

Algebra suffers from a motivational problem. Examples of the useless artificiality of most algebraic problems abound in every textbook ever written on the subject. Here, for example, is a problem from Girolamo Cardano's book *Ars magna* (1545):

Two men go into business together and have an unknown capital.
Their gain is equal to the cube of the tenth part of their capital.
If they had made three ducats less, they would have gained an amount exactly equal to their capital. What was the capital and their profit? [Quoted by Pesic, (2003), pp. 30–31]

If reading this problem makes you want to suggest, "Let's just ask them what their capital and profit were," you are to be congratulated on your astuteness. The second statement of the problem, in particular, marks the entire scenario as an airy flight of fancy. Where in the world would anyone get this kind of information? What data banks is it kept in? How could anyone know this relationship between capital and profit without knowing what the capital and profit were? One of the hardest questions to answer in teaching either algebra or its history is "What is it *for*?" Although some interesting *algebra problems* can be generated from geometric figures, it is not clear that these problems are interesting *as geometry*. Reading the famous treatises on algebra, we might conclude that it is pursued for amusement by people who like puzzles.¹ That answer is not very satisfying, however, and we shall be on the alert for better motivations as we study the relevant documents.

1. Egypt

Although arithmetic and geometry fill up most of the Egyptian papyri, there are some problems in them that can be considered algebra. These problems tend to be what we now classify as linear problems, since they involve the implicit use of direct proportion. The concept of proportion is the key to the problems based on the "rule of false position." Problem 24 of the Ahmose Papyrus, for example, asks for the quantity that yields 19 when its seventh part is added to it. The author notes that if the quantity were 7 (the "false [sup]position"), it would yield 8 when its seventh part is added to it. Therefore, the correct quantity will be obtained by performing the same operations on the number 7 that yield 19 when performed on the number 8. The Egyptian format for such computations is well adapted for handling problems of this sort. The key to the solution seems to be, implicitly, the

¹ In one episode of a popular American situation comedy series during the 1980s, a young police-woman was working undercover, pretending to be a high-school student. While studying algebra with a classmate, she encountered a problem akin to the following. "Johnny is one-third as old as his father; in 15 years he will be half as old. How old are Johnny and his father?" Her response—a triumph of common sense over a mindless educational system—was: "Do we know these people?"

notion that multiplication is distributive over addition (another way of saying that proportions are preserved). But of course, since multiplication was thought of in a peculiar way in Egyptian culture, the algebraic reasoning was very likely as follows: Such-and-such operations applied to 8 will yield 19. If I first add the seventh part of 7 to 7, I will get 8 as a result. If I then perform those operations on 8, I will get 19. *Therefore*, if I first perform those operations on 7, and then add the seventh part of the result to itself, I will also get 19.

The computation is carried out by the standard Egyptian method. First find the operations that must be performed on 8 in order to yield 19:

1	8
2	16 *
$\bar{2}$	4
$\bar{4}$	2 *
$\bar{8}$	1 *
2 $\bar{4}$ $\bar{8}$	19 Result .

Next, perform these operations on 7:

1	7
2	14 *
$\bar{2}$	3 $\bar{2}$
$\bar{4}$	1 $\bar{2}$ $\bar{4}$ *
$\bar{8}$	$\bar{2}$ $\bar{4}$ $\bar{8}$ *
2 $\bar{4}$ $\bar{8}$	16 $\bar{2}$ $\bar{8}$ Result .

This is the answer. The scribe seems quite confident of the answer and does not carry out the computation needed to verify that it works.

The Egyptian scribes were capable of performing operations more complicated than mere proportion. They could take the square root of a number, which they called a *corner*. The Berlin Papyrus 6619, contains the following problem (Gillings, 1972, p. 161):

The area of a square of 100 is equal to that of two smaller squares.
The side of one is $\bar{2}$ $\bar{4}$ the side of the other. Let me know the sides
of the two unknown squares.

Here we are asking for two quantities given their ratio ($\frac{3}{4}$) and the sum of their squares (100). The scribe assumes that one of the squares has side 1 and the other has side $\bar{2}$ $\bar{4}$. Since the resulting total area is $1 \bar{2} \bar{16}$, the square root of this quantity is taken ($1 \bar{4}$), yielding the side of a square equal to the sum of these two given squares. This side is then multiplied by the correct proportionality factor so as to yield 10 (the square root of 100). That is, the number 10 is divided by $1 \bar{4}$, giving 8 as the side of the larger square and hence 6 as the side of the smaller square. This example, incidentally, was cited by van der Waerden as evidence of early knowledge of the Pythagorean theorem in Egypt.

2. Mesopotamia

If we interpret Mesopotamian algebra in our own terms, we can credit the mathematicians of that culture with knowing how to solve some systems of two linear equations in two unknowns, any quadratic equation having at least one real positive root, some systems of two equations where one of the equations is linear and the other quadratic, and a potentially complete set of cubic equations. Of course, it must be remembered that these people were solving *problems*, not *equations*. They did not have any classification of equations in which some forms were solvable and others not. What they knew was that they could find certain numbers from certain data.

2.1. Linear and quadratic problems. As mentioned in Section 4 of Chapter 6, the Mesopotamian approach to algebraic problems was to associate with every pair of numbers another pair: their average and their *semidifference*. These linear problems arise frequently as a subroutine in the solution of more complex problems involving squares and products of unknowns. In Mesopotamia, quadratic equations occur most often as problems in two unknown quantities, usually the length and width of a rectangle. The Mesopotamian mathematicians were able to reduce a large number of problems to the form in which the sum and product or the difference and product of two unknown numbers are given. We shall consider an example that has been written about by many authors. It occurs on a tablet from the Louvre in Paris, known as AO 8862.²

A loose translation of the text of this tablet, made from Neugebauer's German translation, reads as follows:

I have multiplied the length and width so as to make the area. Then I added to the area the amount by which the length exceeds the width, obtaining 3,3. Then I added the length and width together, obtaining 27. What are the length, width, and area?

27	3,3 the sums
15	length
	3,0 area
12	width

You proceed as follows:

Add the sum (27) of the length and width to 3,3. You thereby obtain 3,30. Next add 2 to 27, getting 29. You then divide 29 in half, getting 14;30. The square of 14;30 is 3,30;15. You subtract 3,30 from 3,30;15, leaving the difference of 0;15. The square root of 0;15 is 0;30. Adding 0;30 to the original 14;30 gives 15, which is the length. Subtracting 0;30 from 14;30 gives 14 as width. You then subtract 2, which was added to the 27, from 14, giving 12 as the final width.

The author continues, verifying that these numbers do indeed solve the problem. This text requires some commentary, since it is baffling at first. Knowing the general approach of the Mesopotamian mathematicians to problems of this sort, one can understand the reason for dividing 29 in half (so as to get the average of two numbers) and the reason for subtracting 3,30 from the square of 14;30 (the

² AO stands for *Antiquités Orientales*.

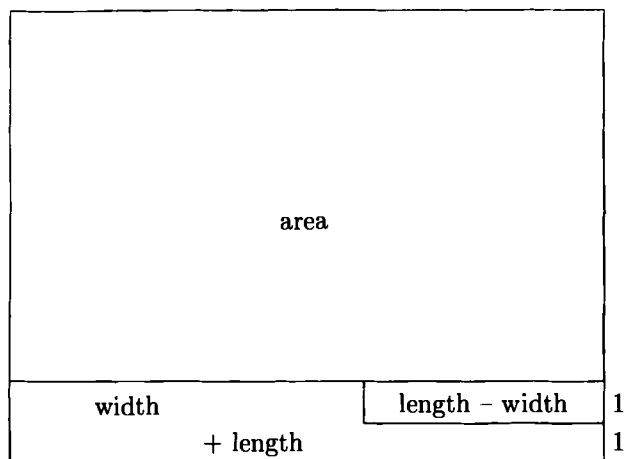


FIGURE 1. Reduction of a problem to standard form.

difference between the square of the average and the product will be the square of the semidifference of the two numbers whose sum is 29 and whose product is 3,30, that is, 210). What is not clear is the following: Why add 27 to the number 3,3 in the first place, and why add 2 to 27? Possibly the answer is contained in Fig. 1, which shows that adding the difference between length and width to the area amounts to gluing a smaller rectangle of unit width onto a larger rectangle. Then adding the sum of length and width amounts to gluing a gnomon onto the resulting figure in order to complete a rectangle two units wider than the original. Finding the dimensions of that rectangle from its perimeter and area is the standard technique of solving a quadratic equation, and that is what the author does.

The tablet AO 6670, discussed by van der Waerden (1963, pp. 73–74) contains a rare explanation of the procedure for solving a problem that involves two unknowns and two conditions, given in abstract terms without specific numbers. Unfortunately, the explanation is very difficult to understand. The statement of the problem is taken directly from Neugebauer's translation: *Length and width as much as area; let them be equal*. Thereafter, the translation given by van der Waerden, due to François Thureau-Dangin (1872–1944), goes as follows:

The product you take twice. From this you subtract 1. You form the reciprocal. With the product that you have taken you multiply, and the width it gives you.

Van der Waerden asserts that the formula $y = (1/(x - 1)) \cdot x$ is “stated in the text” of Thureau-Dangin's translation. If so, it must have been stated in a place not quoted by van der Waerden, since x is not a “product” here, nor is it taken twice. Van der Waerden also notes that according to Evert Marie Bruins (1909–1990), the phrase “length and width” does not mean the *sum* of length and width. Van der Waerden says that “the meaning of the words has to be determined in relation to the mathematical content.” The last two sentences in the description tell how to determine the width once the length has been found. That is, you take

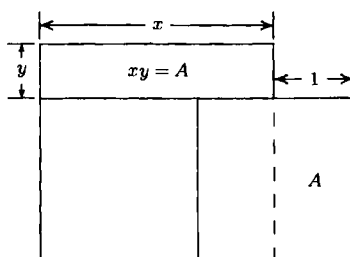


FIGURE 2. A scenario that may “fit” a text from cuneiform tablet AO 6670.

reciprocal of the length and multiply it by the product of length and width, which must be given in the problem as the area. The mystery is then pushed into the first two instructions. What product is being “taken twice”? Does taking a product twice mean multiplying by 2, or does it mean cubing? Why is the number 1 being subtracted? Perhaps we should go back to the original statement and ask whether “as much as area” implies an equation, or whether it simply means that length and width *form* an area. What does the word *them* refer to in the statement, “Let them be equal”? Is it the length and the width, or some combination of them and the area? Without knowing the original language and seeing the original text, we cannot do anything except suggest possible meanings, based on what is mathematically correct, to those who do know the language.

We can get a geometric problem that fits this description by considering Fig. 2, where two equal squares have been placed side by side and a rectangle of unit length, shown by the dashed line, has been removed from the end. If the problem is to construct a rectangle on the remaining base equal to the part that was cut off, we have conditions that satisfy the instructions in the problem. That is, the length x of the base of the new rectangle is obtained numerically by subtracting 1 from twice the given area. This scenario is fanciful, however, and is not seriously proposed as an explanation of the text. Another scenario that “explains” the text can be found in Problem 13.4.

2.2. Higher-degree problems. Cuneiform tablets have been found that give the sum of the square and the cube of an integer for many values of the integer. These tablets may have been used for finding the numbers to which this operation was applied in order to obtain a given number. In our terms these tablets make it possible to solve the equation $x^3 + x^2 = a$, a very difficult problem indeed. In fact, given a complete table of $x^3 + x^2$, one can solve every cubic equation $ay^3 + by^2 + cy = d$, where b and c are nonnegative numbers and a and d are positive. (See Problem 13.5.)

Neugebauer (1935, p. 99; 1952, p. 43) reports that the Mesopotamian mathematicians moved beyond algebra proper and investigated the laws of exponents, compiling tables of successive powers of numbers and determining the power to which one number must be raised in order to yield another. Such problems occur in a commercial context, involving compound interest. For example, the tablet AO 6484 gives the sum of the powers of 2 from 0 to 9 as the last term plus one less than the last term, and the sum of the squares of the first segment of integers as the sum of the same integers multiplied by the sum of $\frac{1}{3}$ and $\frac{2}{3}$ of the last term.

This recipe is equivalent to the modern formula for the sum of the squares of the first n integers.

3. India

Problems leading to algebra can be found in the *Sulva Sutras* and the *Bakshali Manuscript*, mentioned in Section 2 of Chapter 2. Since we have already discussed the Diophantine-type equations that result from the altar-construction problems in the *Sulva Sutras*, we confine ourselves here to a few problems that lead to linear Diophantine equations, determinate quadratic equations, and the summation of progressions.

3.1. Jaina algebra. According to Srinivasiengar (1967, p. 25), by the year 300 BCE Jaina mathematicians understood certain cases of the laws of exponents. They could make sense of an expression like $a^{m/2^n}$, interpreting it as extracting the square root n times and then raising the result to the power m . The notation used was of course not ours. The power $\frac{3}{4}$, for example, was described as "the cube of the second square root." That the laws of exponents were understood for these special values is attested by such statements as "the second square root multiplied by the third square root, or the cube of the third square root," indicating an understanding of the equality

$$a^{1/4}a^{1/8} = a^{3/8}.$$

3.2. The Bakshali Manuscript. The birchbark manuscript discovered in the village of Bakshali, near Peshawar, in 1881 uses the symbol \ominus to denote an unknown quantity. One of the problems in the manuscript is written as follows, using modern number symbols and a transliteration of the Sanskrit into the Latin alphabet:

$$\begin{array}{ccccccc} \ominus & 5 & & \ominus & & \ominus & 7 + & m\bar{u} & \ominus \\ 1 & 1 & yu & m\bar{u} & 1 & sa & 1 & 1 & 1 \end{array}.$$

This symbolism can be translated as, "a certain thing is increased by 5 and the square root is taken, giving [another] thing; and the thing is decreased by 7 and the square root is taken, giving [yet another] thing." In other words, we are looking for a number x such that $x + 5$ and $x - 7$ are both perfect squares. This problem is remarkably like certain problems in Diophantus. For example, Problem 11 of Book 2 of Diophantus is to add the same number to two given numbers so as to make each of them a square. If the two given numbers are 5 and -7 , this is *exactly* the problem stated here; Diophantus, however, did not use negative numbers.

The Bakshali Manuscript also contains problems in linear equations, of the sort that have had a long history in elementary mathematics texts. For example, three persons possess seven thoroughbred horses, nine draft horses, and 10 camels respectively. Each gives one animal to each of the others. The three are then equally wealthy. Find the (relative) prices of the three animals. Before leaping blindly into the set of two linear equations in three unknowns that this problem prescribes, we should take time to note that the problem can be solved by imagining the experiment actually performed. Suppose that these donations have been made and the three people are now equally wealthy. They will remain equally wealthy if each gives away one thoroughbred horse, one draft horse, and one camel. It follows that four thoroughbred horses, six draft horses, and seven camels are all of equal value. The problem has thereby been solved, and no actual algebra has been

performed. Srinivasiengar (1967, p. 39) gives the solution using symbols for the unknown values of the animals, but does not assert that the solution is given this way in the manuscript itself.

4. China

With the exception of the the *Zhou Bi Suan Jing*, which is mostly about geometry and astronomy, algebra forms a major part of early Chinese mathematical works. The difficulty with finding early examples of problems leading to algebra is that the earliest document after the *Zhou Bi Suan Jing*, the *Jiu Zhang Suanshu*, contains not only many problems leading to systems of linear equations but also a sophisticated method of solving these equations, fully equivalent to what we now call Gaussian elimination (row reduction) of matrices and known as *fang cheng* or the *rectangular algorithm*. Li and Du (1987, pp. 46–47) discuss one example involving the yield of three different kinds of grain, in which a matrix is triangularized so that the solution can be obtained by working from bottom to top. Our discussion of this technique, like the discussion of quadratic equations, is reserved for Chapter 14.

4.1. The *Jiu Zhang Suanshu*. In Chapter 6 of the *Jiu Zhang Suanshu* we find some typical problems leading to one linear equation in one unknown. This type of problem can be solved using algebra, but does not necessarily *require* algebraic reasoning to solve, since the answer lies very close to the surface. For example (Mikami, 1913, p. 16), if a fast walker goes 100 paces in the time required for a slow walker to go 60 paces, and the slower walker has a head start of 100 paces, how many paces will be required for the fast walker to overtake the slow one? The instruction given is to multiply the head start by the faster speed and divide by the difference in speeds. That will obviously give the number of paces taken by the faster runner. The author says nothing about the number of paces that will be taken by the slower runner, but he probably noticed that that number could be obtained in two ways: by subtracting 100 (the head start) or by multiplying by the slower speed instead of the faster. This equivalence, if noticed, would give some insight into manipulating expressions for numbers.

Chapter 7 contains the kind of excess–deficiency problems discussed in Section 2 of Chapter 6. The solutions are described in some detail, so that we can judge the extent to which they are to be considered algebra. For example, an unknown number of people are buying hens. If each gives nine (units of money), there will be a surplus of 11 units. If each gives six, there will be a deficit of 16. The instructions for solution are to arrange the data in a rectangle, cross-multiply, and add the products. In other words, form the number $9 \cdot 16 + 6 \cdot 11 = 210$. If this number is divided by the difference $9 - 6$, the result, 70, represents the total price to be paid. Adding the surplus and deficit gives 27, and when this is divided by $9 - 6$, we get 9, the number of people buying. This solution is far too sophisticated and general to be an early method aimed at one specific problem. It is algebra proper.

4.2. The *Suanshu Shu*. Li and Du (1987, pp. 56–57) describe a set of bamboo strips discovered in 1983–1984 in three tombs from the western Han Dynasty containing a *Suanshu Shu* (*Arithmetical Book*) and dated no later than the first half of the second century BCE. This work contains instructions on the performance of arithmetical operations and some applications that border on algebra. For example, one problem is to find the width of a field whose area is 1 *mu* (240 square *bu*)

and whose length is $1\frac{1}{2}$ *bu*. This problem amounts to one linear equation in one unknown. Dividing the area by the length yields 160 *bu* as the answer. The whole difficulty of the problem lies in the complicated rules for dividing by a fraction.

4.3. The *Sun Zi Suan Jing*. A large number of problems leading to algebra are considered in the *Sun Zi Suan Jing*. Some of these are the kind of excess and deficit problems already discussed. Others involve arithmetic and geometric progressions and are solved by clever numerical reasoning. As an example of an arithmetic progression, Problem 25 of Chapter 2 of the *Sun Zi Suan Jing* discusses the distribution of 60 tangerines among five noblemen of different ranks in such a way that each will receive three more than the one below him. Sun Zi says first to give the lowest-ranking nobleman three, then six to the next-higher rank, and so on, until the fifth person gets 15. That accounts for $3 + 6 + 9 + 12 + 15 = 45$ tangerines and leaves fifteen more to be divided equally among the five. Thus the numbers given out are 6, 9, 12, 15, and 18.

4.4. Zhang Qiujian. To the fifth-century mathematician Zhang Qiujian (ca. 430–490) we owe one of the most famous and long-lasting problems in the history of algebra. It goes by the name of the Hundred Fowls Problem, and reads as follows: Roosters cost 5 qian each, hens 3 qian each, and three baby chicks cost 1 qian. If 100 fowls are bought for 100 qian, how many roosters, hens, and chicks were bought? The answer is not unique, but Zhang gives all the physically possible solutions: (4, 18, 78), (8, 11, 81), and (12, 4, 84). Probably this answer was obtained by enumeration. Given that one is to buy at least one of each type of chicken, at most 19 roosters can be bought. Zhang observed that the number of roosters must increase in increments of 4, the number of hens must decrease in increments of 7, and the number of baby chicks must increase in increments of 3. That is because $4 - 7 + 3 = 0$ and $4 \cdot 5 - 7 \cdot 3 + 3 \cdot \frac{1}{3} = 0$.

According to Mikami (1913, p. 41), three other “hardy perennials” of algebra can be traced to Zhan Qiujian’s treatise. One involves arithmetic progression. A weaver produces 5 feet of fabric on the first day, and the output diminishes (by the same amount) each day, until only 1 foot is produced on the thirtieth day. What was the total production? The recipe for the answer is to add the amounts on the first and last days, divide by 2, and multiply by the number of days.

The second is a rate problem of the type found in the *Jiu Zhang Suanshu*. A horse thief rode 37 miles before his theft was discovered. The owner then pursued him for 145 miles and narrowed the distance between them to 23 miles, but gave up at that point and returned home. If he had continued the pursuit, how many more miles would he have had to ride to catch the thief? Here we have the case of one person traveling 145 miles in the same time required for the other to travel 131 miles, and the other person having a 23-mile head start. Following the formula given in the *Jiu Zhang Suanshu*, Zhang Qiujian gives the answer as $145 \times 23 \div 14$.

Finally, we have another rate problem: If seven men construct $12\frac{1}{2}$ bows in nine days, how many days will be required for 17 men to construct 15 bows?

All these problems can be solved by *reasoning about numbers* without necessarily writing down any equations. But they are definitely proto-algebra in that they require thinking about performing operations on abstract, unspecified numbers.

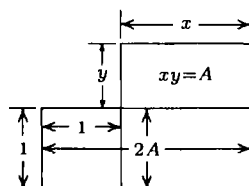


FIGURE 3. Another scenario to “fit” a text on cuneiform tablet AO 6670.

Questions and problems

13.1. What do the two problems of recovering two numbers from their sum and product or from their difference and product have to do with quadratic equations as we understand them today? Can we conclude that the Mesopotamians “did algebra”?

13.2. You can verify that the solution of the problem from tablet AO 8862 (15 and 12) given by the author is not the only possible one. The numbers 14 and 13 will also satisfy the conditions of the problem. Why didn’t the author give this solution?

13.3. Of what practical value are the problems we have called “algebra”? Taking just the quadratic equation as an example, the data can be construed as the area and the semiperimeter of a rectangle and the solutions as the sides of the rectangle. What need, if any, could there be for solving such a problem? Where are you ever given the perimeter and area of a rectangle and asked to find its shape?

13.4. Figure 3 gives a scenario that can be fit to the data in AO 6670. Given a square 1 unit on a side, in the right angle opposite one of its corners construct a rectangle of prescribed area A that will be one-third of the completed gnomon. Explain how the figure fits the statement of the problem. (As in Section 2, this scenario is *not* being proposed as a serious explanation of the text.)

13.5. Given a cubic equation

$$ax^3 + bx^2 + cx = d,$$

where all coefficients are assumed positive, let $A = d + bc/(3a) - 2b^3/(9a^2)$, $B = b^2/(3a) - c$, and $t = 3aA/(3aBx - bB)$, that is, $x = A/(Bt) + b/(3a)$. Show that in terms of these new parameters, this equation is

$$t^3 + t^2 = \frac{aA^2}{B^3}.$$

It could therefore be solved numerically by consulting a table of values of $t^3 + t^2$. [Again a caution: The fact that such a table exists and could be used this way does not imply that it *was* used this way, any more than the fact that a saucer can be used to hold paper clips implies that it was designed for that purpose.]

13.6. Considering the origin of algebra in the mathematical traditions we have studied, do you find a point in their development at which mathematics ceases to be a disjointed collection of techniques and becomes systematic? What criteria would you use for defining such a point, and where would you place it in the mathematics of Egypt, Mesopotamia, Greece, China, and India?

CHAPTER 14

Equations and Algorithms

In this chapter we take up the history of algebra from the point at which equations appear explicitly and carry it forward along two parallel lines. In one line of development the emphasis is on finding numerical approximations to the roots of an equation. In the second line the emphasis is on finding an algorithm involving only the four operations of arithmetic and the extraction of roots that will yield the solution. The second line of development reached its highest point of achievement in sixteenth-century Italy, with the arithmetical solution of equations of degree 4. That is the point at which the present chapter ends. Standing somewhat to one side of both lines of evolution was the work of Diophantus, which contains a mixture of topics that now form part of number theory and algebra.

1. The *Arithmetica* of Diophantus

The work of Diophantus of Alexandria occupies a special place in the history of algebra. To judge it, one should know something of its predecessors and its influence. Unfortunately, information about either of these is difficult to come by. The Greek versions of the treatise, of which there are 28 manuscripts, according to Sesiano (1982, p. 14), all date to the thirteenth century. Among the predecessors of Diophantus, we can count Heron of Alexandria and one very obscure Thymaridas, who showed how to solve a particular set of linear equations, the *epanthēma* (blossom) of Thymaridas. Because the work of Diophantus is so different from the Pythagorean style found in Euclid and his immediate successors, the origins of his work have been traced to other cultures, notably Egypt and Mesopotamia. The historian of mathematics Paul Tannery (1843–1904) printed an edition of Diophantus' work and included a fragment supposedly written by the eleventh-century writer Michael Psellus (1018–ca. 1078), which stated that “As for this *Egyptian* method, while Diophantus developed it in more detail, . . .” It was on this basis, identifying Anatolius with a third-century Bishop of Laodicea originally from Alexandria, that Tannery assigned Diophantus to the third century. Neugebauer (1952, p. 80) distinguishes two threads in Hellenistic mathematics, one in the logical tradition of Euclid, the other having roots in the Babylonian and Egyptian procedures and says that, “the writings of Heron and Diophantus . . . form part of this oriental tradition which can be followed into the Middle Ages both in the Arabic and in the western world.” Neugebauer sees Diophantus as reflecting an earlier type of mathematics practiced in Greece alongside the Pythagorean mathematics and temporarily eclipsed by the Euclidean school. As he says (1952, p. 142):

It seems to me characteristic, however, that Archytas of Tarentum could make the statement that not geometry but arithmetic alone

could provide satisfactory proofs. If this was the opinion of a leading mathematician of the generation just preceding the birth of the axiomatic method, then it is rather obvious that early Greek mathematics cannot have been very different from the Heronic Diophantine type.

1.1. Diophantine equations. An equation containing two or more unknowns for which only rational (or more often, integer) solutions are sought is nowadays called a *Diophantine equation*. Diophantus wrote a treatise commonly known under the somewhat misleading name *Arithmetica*. As mentioned previously, the six books of this treatise that have been known for some centuries may now be supplemented by parts of four other books, discovered in 1968, but that is not certain. The title itself is of some interest. Its suffix *-tica* has come into English from Greek in a large number of words such as *logistics*, *mathematics*, and *gymnastics*. It has a sense of how-to, that is, the techniques involved in using numbers (*arithmoí*) or reasoning (*lógoi*) or learning (*mathēmata*) or physical training (*gýmnasis*).¹ The plural *-s* on these English words, even though they are now regarded as singular, reflects the fact that these words were originally intended to be plural—the neuter plural form of the corresponding adjectives *arithmētikós* (adept with number), *logistikós* (skilled in calculating), *mathēmatikós* (disposed to learn), and *gymnastikós* (skilled in bodily exercise), but which evolved into a feminine singular form. The Greek title of the work is *Diophántou Alexándreōs Arithmētikôn*, meaning [The Books] of *Arithmetics of Diophantus of Alexandria*.

1.2. General characteristics of the *Arithmetica*. In contrast to other ancient works containing problems that lead to algebra, the problems that require algebraic techniques in the *Arithmetica* all involve purely numerical relations. They are not problems about *things* that have been counted or measured. They are about *counting itself*. The work begins with a note to one Dionysius, whom the author characterizes as “cager to learn” how to solve problems in arithmetic.² In a number of ways Diophantus seems to be doing something that resembles the algebra taught nowadays. In particular, he has a symbol for an unknown or abstract number that is to be found in a problem, and he appears to know what an equation is, although he doesn’t exactly use the word *equation*.

Diophantus began by introducing a symbol for a constant unit $\overset{\circ}{M}$, from *monás* (*μονάς*), along with a symbol for an unknown number ς , conjectured to be an abbreviation of the first two letters of the Greek word for number: *arithmós* (*ἀριθμός*). For the square of an unknown he used Δ^v , the first two letters of *dýnamis* (*δύναμις*), meaning *power*. For its cube he used K^v , the first two letters of *kýbos* (*κύβος*), meaning *cube*. He then combined these letters to get fourth ($\Delta^v\Delta$), fifth (ΔK^v), and sixth (K^vK) powers. For the reciprocals of these powers of the unknown he invented names by adjoining the suffix *-ton* (*-τον*) to the names of the corresponding powers. These various powers of the unknown were called *eída* (*εἶδα*), meaning *species*. Diophantus’ system for writing down the equivalent of a polynomial in the unknown consisted of writing down these symbols in order

¹ From the root *gymnós*, meaning *naked*.

² One of the reasons that Tannery assigned Dionysius to the third century was that this date made it easy to imagine that Dionysius was the man appointed Bishop of Alexandria in 247—not that Dionysius (Dennis) was exactly a rare name in those days.

to indicate addition, each term followed by the corresponding number symbol (for which the Greeks used their alphabet). Terms to be added were placed first, separated by a pitchfork (\pitchfork) from those to be subtracted. Heath conjectured that this pitchfork symbol is a condensation of the letters lambda and iota, the first two letters of a Greek root meaning *less* or *leave*. Thus what we would call the expression $2x^4 - x^3 - 3x^2 + 4x + 2$ would be written $\Delta^v \Delta \bar{\beta} \zeta \bar{\delta} \overset{\circ}{M} \bar{\beta} \pitchfork K^v \bar{\alpha} \Delta^v \bar{\gamma}$.

Diophantus' use of symbolism is rather sparing by modern standards; he often uses words where we would use symbolic manipulation. For this reason his algebra was described by the nineteenth-century German historian of mathematics Nesselmann as a transitional "syncopated" phase between the earliest "rhetorical" algebra, in which everything is written out in words, and the modern "symbolic" algebra.³

1.3. Determinate problems. The determinate problems in the *Arithmetica* require that one or more unknown numbers be found from conditions that we would nowadays write as systems of linear or quadratic equations. The 39 problems of Book 1 and the first ten problems of Book 2 are of these types. Some of these problems have a unique solution. For example, Problem 7 of Book 1 is: *From a given unknown number subtract two given numbers so that the remainders have a given ratio*. In our terms, this condition says

$$x - a = m(x - b),$$

where x is unknown, a and b are the given numbers, and m is the given ratio. Since it is obvious that $m \geq 1$ if all quantities are positive and $a \leq b$, Diophantus has no need to state this restriction.

Similarly, Problem 15 of Book 1 asks for two numbers (x and y , we would say) such that for given numbers a and b the ratios $x + a : y - a$ and $y + b : x - b$ are equal to two given ratios r and s .

The symbolic notation of Diophantus extended only as far as the unknown and representations of sums, products, and differences. He had no way of forming mathematical expressions containing the phrases "a given number" (a and b above) and "a given ratio" (r and s above). As a result, he could explain his methods of solution only by using a particular example, in the present case taking $a = 30$, $r = 2$, $b = 50$, $s = 3$. He then assumed that $y = \zeta + 30$ and $x = 2\zeta - 30$, so that the first equation was satisfied automatically and the second became $\zeta + 80 = 3(2\zeta - 80)$. Here it is very easy to recognize the explicit manipulation of formal expressions, leading to the discovery of the unknown number. This manipulation of expressions is characteristic of algebraic technique.

Some of the problems that are determinate from our point of view may have no positive rational solutions for certain data, and in such cases Diophantus requires a restriction on the data so that positive rational solutions will exist. For example, Problem 8 of Book 1 is *to add the same (unknown) number to two given numbers so that the sums have a given ratio*. This problem amounts to the equation

$$x + a = m(x + b).$$

³ Nesselmann is quoted by Jacob Klein (1934-36, p. 146). In the author's opinion, there is not much for Diophantus to be transitional between, since little is known of his algebraic predecessors, and later algebraists wrote everything out in words. Jacob Klein seems to share these reservations.

If $x > 0$ and $a > b$, then $1 < m = (x + a)/(x + b) < a/b$. That is, the given ratio must be larger than 1 and less than the ratio of the larger number to the smaller.

1.4. The significance of the *Arithmetica*. The existence of Arabic manuscripts of Diophantus' treatise shows that his work was known to the Muslim mathematicians of the Middle Ages. Sesiano (1982, pp. 9–20) discusses the extent to which a number of Islamic and Byzantine mathematicians were influenced by his work or commented on it. He comments (p. 9) that, "There is nothing to suggest that the Egyptian Abu-Kamil had any direct (or even indirect) knowledge of Diophantus' *Arithmetica*, although the problems in his *Algebra* dealing with indeterminate analysis are perfectly Diophantine in form and the basic methods are attested to in the *Arithmetica*." In contrast, the Diophantine connection is clear in the case of the eleventh-century mathematician al-Karkhi, (also known as al-Karaji, 953–1029), whose *Fakhri* has many points of contact with Diophantus. Tracing the influence of Diophantus, however, is more difficult. Jacob Klein (1934–36, p. 5), citing nineteenth-century work of Tannery and others, says that "the special influence of the *Arithmetic* of Diophantus on the content, but even more so on the form, of this Arabic science is unmistakable – if not in the *Liber Algorismi* of Al-Khowarizmi himself, at any rate from the tenth century on." In a treatise on algebra published in the late sixteenth century, the engineer mathematician Rafael Bombelli stated that, although it had been agreed up to his time that algebra was an invention of the Muslims, he was convinced, after reading the work of Diophantus, that the invention should be ascribed to the latter.

At the very least, Diophantus used equations and developed a symbolism for handling algebraic expressions, and that, in the long run, was an important innovation. As two prominent Russian historians of science say:

Diophantus was the first to deduce that it was possible to formulate the conditions of a problem as equations or systems of equations; as a matter of fact, before Diophantus, there were no equations at all, either determinate or indeterminate. Problems were studied that we can now reduce to equations, but nothing more than that. [Bashmakova and Smirnova 1997, p. 132]

1.5. The view of Jacob Klein. In several places in Part 2 and in the present part we have used without comment the "standard view" among historians of a contrast between *logistikē* and *arithmētikē* in the science and philosophy of ancient Greece, *logistikē* being counting or computation and *arithmētikē* being the study of the theoretical properties of numbers. A different point of view is contained in the extended essay by Jacob Klein (1934–36). Klein maintains that even the word *arithmós* itself has been misinterpreted, that Euclid and Diophantus did not have in mind cardinal numbers in the abstract, but used the word *arithmós* to mean a set or collection. As he says (p. 7), "*arithmós* never means anything other than 'a definite number of definite objects.'" He goes on to say (p. 19) that for Plato "arithmetical" is, accordingly, not 'number theory,' but first and foremost the art of correct counting."⁴ In particular, Klein denies that Euclid was thinking about

⁴ Such may well be the case. If so, that is unfortunate for Plato's reputation. Neugebauer (1952, p. 146) offers the opinion that "Plato's role has been widely exaggerated. His own direct contributions to mathematical knowledge were obviously nil... The often adopted notion that Plato 'directed' research fortunately is not borne out by the facts."

numbers in the abstract and illustrating them geometrically with lines. It does seem strange that Euclid clings to what appears to be a completely unnecessary geometric representation of a number. According to Klein, the mystery is solved if we recognize that specific numbers were always intended, even though an abstract symbol (a letter or two letters) was used for them. Klein (p. 124) cites Tannery in arguing that these letters did not represent general, unspecified numbers, because they were not amenable to being operated on.

2. China

The development of algebra in China began early and continued for many centuries. The aim was to find numerical approximate solutions to equations, and the Chinese mathematicians were not intimidated by equations of high degree.

2.1. Linear equations. We have already mentioned the Chinese technique of solving simultaneous linear equations and pointed out its similarity to modern matrix techniques. Examples of this method are found in the *Jiu Zhang Suanshu* (Mikami, 1913, pp. 18–22; Li and Du, 1987, pp. 46–49). Here is one example of the technique.

There are three kinds of [wheat]. The grains contained in two, three and four bundles, respectively, of these three classes of [wheat], are not sufficient to make a whole measure. If however we add to them one bundle of the second, third, and first classes, respectively, then the grains would become one full measure in each case. How many measures of grain does then each one bundle of the different classes contain?

The following counting-board arrangement is given for this problem.

1		2	1st class
	3	1	2nd class
4	1		3rd class
1	1	1	measures

Here the columns from right to left represent the three samples of wheat. Thus the right-hand column represents 2 bundles of the first class of wheat, to which one bundle of the second class has been added. The bottom row gives the result in each case: 1 measure of wheat. The word problem might be clearer if the final result is thought of as the result of threshing the raw wheat to produce pure grain. We can easily, and without much distortion in the procedure followed by the author, write down this counting board as a matrix and solve the resulting system of three equations in three unknowns. The author gives the solution: A bundle of the first type of wheat contains $\frac{9}{25}$ measure, a bundle of the second $\frac{7}{25}$ measure, and a bundle of the third $\frac{4}{25}$ measure.

2.2. Quadratic equations. The last chapter of the *Jiu Zhang Suanshu*, which involves right triangles, contains problems that lead to linear and quadratic equations. For example (Mikami, 1913, p. 24), there are several problems involving a town enclosed by a square wall with a gate in the center of each side. In some cases the problem asks at what distance from the south gate a tree a given distance east of the east gate will first be visible. The data are the side s of the square and the

distance d of the tree from the gate. For that kind of data, the problem is the linear equation $(x + s/2)/(s/2 + d) = s/(2d)$. When the side of the town is the unknown, a quadratic equation results, as in one case, in which it is asserted that the tree is 20 paces north of the north gate and is just visible to a person who walks 14 paces south of the south gate, then 1775 paces west. This problem proposes a quadratic equation as a problem to be solved for a single unknown number, in contrast to the occurrence of quadratic equations in Mesopotamia, where they amount to finding two numbers given their sum and product. Since the Chinese technique of solving equations numerically is practically independent of degree, we shall not bother to discuss the techniques of solving quadratic equations separately.

2.3. Cubic equations. Cubic equations first appear in Chinese mathematics (Li and Du, 1987, p. 100; Mikami, 1913, p. 53) in the seventh-century work *Xugu Suanjing* (*Continuation of Ancient Mathematics*) by Wang Xiaotong. This work contains some intricate problems associated with right triangles. For example, compute the length of a leg of a right triangle given that the product of the other leg and the hypotenuse is $1337\frac{1}{20}$ and the difference between the hypotenuse and the leg is $1\frac{1}{10}$.⁵ Obviously, the data are perfectly general for a product P and a difference D . Wang Xiaotong gives a general description of the result of eliminating the hypotenuse and the other leg that amounts to the equation

$$x^3 + \frac{5D}{2}x^2 + 2D^2x = \frac{P^2}{2D} - \frac{D^3}{2}.$$

In this particular case the equation is

$$x^3 + \frac{1}{4}x^2 + \frac{1}{50}x - 8938513\frac{64}{125} = 0.$$

He then says to compute the root (which he gives as $92\frac{2}{5}$) "according to the rule of the cubic root extraction." Li and Du (1987, pp. 118–119) report that the eleventh-century mathematician Jia Xian developed the following method for extracting the cube root. This method generalizes from the case $x^3 = N$ to the general cubic equation quite easily, as we shall see.

The computation is arranged in rows (or columns) of five elements. We shall use columns for convenience. The top entry is always the current approximation a to the cube root, the bottom entry is always 1. The entries in the next-to-bottom and middle rows are obtained successively by multiplying the entry that was just below at the preceding stage by the adjustment and adding to the entry that was in the same row at the preceding step. The entry next to the top is obtained the same way, except that the adjustment is subtracted instead of being added. This row always contains the current or adjusted error. The adjustment procedure works first from the bottom to the second row, then from the bottom to the third row, and finally, from the bottom to the fourth row. For example, the first four steps go as follows, assuming a "zeroth" approximation of 0, which is to be improved by an initial guess a :

⁵ Mikami gives $\frac{1}{10}$ as the difference, which is incompatible with the answer given by Wang Xiaotong. I do not know if the mistake is due to Mikami or is in the original.

$$\begin{array}{cccc}
 0 & a & a & a \\
 N & N - a^3 & N - a^3 & N - a^3 \\
 0 \longrightarrow & a^2 & \longrightarrow & 3a^2 \longrightarrow 3a^2 \\
 0 & a & 2a & 3a \\
 1 & 1 & 1 & 1
 \end{array}$$

Next, given any approximation a , the approximation is improved by adding an adjustment b , and the rows are then recomputed, again, first working from the bottom to the second row, then from the bottom to the third row, and finally, from the bottom to the fourth row:

$$\begin{array}{cccc}
 a & a + b & a + b & a + b \\
 N - a^3 & N - (a + b)^3 & N - (a + b)^3 & N - (a + b)^3 \\
 3a^2 & \longrightarrow 3a^2 + 3ab + b^2 & \longrightarrow & 3(a + b)^2 \longrightarrow 3(a + b)^2 \\
 3a & 3a + b & 3a + 2b & 3(a + b) \\
 1 & 1 & 1 & 1
 \end{array}$$

By introduction of a counting board ruled into squares analogous to the registers in a calculator, the procedure could be made completely mechanical. Using an analogous procedure, one can take fifth roots, seventh roots, and so on, with increasingly messy computations, of course. Composite roots can be reduced to prime roots, but since the generalization of this method works so well, there is really no need to do so. The sixth root, for example, can be taken by extracting the square root of the cube root, or it could be extracted directly following this method.

2.4. The numerical solution of equations. The Chinese mathematicians of 800 years ago invented a method of finding numerical approximations of a root of an equation, similar to a method that was rediscovered independently in the nineteenth century in Europe and is commonly called *Horner's method*, in honor of the British school teacher William Horner (1786–1837).⁶ The first appearance of the method is in the work of the thirteenth-century mathematician Qin Jiushao, who applied it in his 1247 treatise *Sushu Jiu Zhang* (*Arithmetic in Nine Chapters*, not to be confused with the *Jiu Zhang Suanshu*).

The connection of this method with the cube root algorithm will be obvious. We illustrate with the case of the cubic equation. Suppose in attempting to solve the cubic equation $px^3 + qx^2 + rx + s = 0$ we have found the first digit (or any approximation) a of the root. We then “reduce” the equation by setting $x = y + a$ and rewriting it. What will the coefficients be when the equation is written in terms of y ? The answer is immediate; the new equation is

$$\begin{array}{ccccccc}
 py^3 & + & 3pay^2 & + & 3pa^2y & + & pa^3 \\
 & + & qy^2 & + & 2qay & + & qa^2 \\
 & & & + & ry & + & ra \\
 & & & & & + & s = 0.
 \end{array}$$

⁶ Besides being known to the Chinese mathematicians 600 years before Horner, this procedure was used by Sharaf al-Tusi (ca. 1135–1213), as discussed in Section 5 below, and was discovered by the Italian mathematician Paolo Ruffini (1765–1822) a few years before Horner published it. In fairness to Horner, it must be said that he applied the method not only to polynomials, but to infinite series representations. To him it was a theorem in calculus, not algebra.

We see that we need to make the following conversion of the coefficients (reading from bottom to top):

$$\begin{array}{rcl} s & pa^3 + qa^2 + ra + s & \\ r & \longrightarrow 3pa^2 + 2qa + r & \\ q & \longrightarrow 3pa + q & \\ p & p & \end{array}.$$

The procedure followed in the cube root algorithm works perfectly. That is, start at the bottom and at each stage, multiply the element below by a and add it to the element in the same row at the preceding stage. Going from bottom all the way to the top gets the top row correct. Then going from bottom to the second row gets the second row correct; and finally going from the bottom to the third row completes the transition:

$$\begin{array}{ccccccc} s & pa^3 + qa^2 + ra + s & pa^3 + qa^2 + ra + s & pa^3 + qa^2 + ra + s & \\ r & \longrightarrow pa^2 + qa + r & \longrightarrow 3pa^2 + 2qa + r & \longrightarrow 3pa^2 + 2qa + r & \\ q & \longrightarrow pa + q & \longrightarrow 2pa + q & \longrightarrow 3pa + q & \\ p & p & p & p & \end{array}.$$

In this context the cube root algorithm itself becomes merely the case $p = 1$, $q = 0 = r$, $s = -N$, with the top row omitted and the subtraction in the second row (now the top row) replaced by addition, since N has been replaced by $-N$. Not only is this algorithm simple to use; it also provides the most efficient and accurate way of computing a polynomial numerically. Before the advent of computer algebra programs, numerical analysis books instructed the student to compute the polynomial $px^3 + qx^2 + rx + s$ at different values of x by the sequence of operations

$$\begin{aligned} p \rightarrow px \rightarrow px + q \rightarrow x(px + q) \rightarrow x(px + q) + r \rightarrow \\ \rightarrow x(x(px + q) + r) \rightarrow x(x(px + q) + r) + s. \end{aligned}$$

This sequence of operations avoids the error that tends to accumulate when large numbers of opposite sign are added.⁷

Wang Xiaotong's reference to the use of cube root extraction for solving his equation seems to suggest that this method was known as early as the seventh century. However, as we have just noted, the earliest explicit record of it seems to be in the treatise of Qin Jiushao, who illustrated it by solving the quartic equation

$$-x^4 + 763200x^2 - 40642560000 = 0.$$

The method of solution gives proof that the Chinese did not think in terms of a quadratic formula. If they had, this equation would have been solved for x^2 using that formula and then x could have been found by taking the square root of any positive root. But Qin Jiushao applied the method described above to get the solution $x = 840$. (He missed the smaller solution $x = 240$.)

The efficiency of this method in finding approximate roots allowed the Chinese to attack equations involving large coefficients and high degrees. Qin Jiushao (Libbrecht, 1973, pp. 134-136) considered the following problem: *Three li north of the wall of a circular town there is a tree. A traveler walking east from the southern gate of the town first sees the tree after walking 9 li. What are the diameter and circumference of the town?*

⁷ In addition, a very simple hand calculator with no memory cells can carry out this sequence of operations without the need to stop entering and write down a partial result.

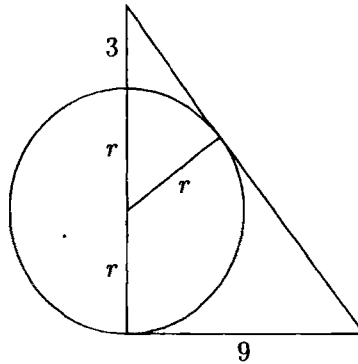


FIGURE 1. A quartic equation problem.

This problem is obviously concocted so as to lead to an equation of higher degree. (The diameter of the town could surely be measured directly from inside, so that it is highly unlikely that anyone would ever need to solve such a problem for a practical purpose.) Representing the diameter of the town as x^2 , Qin Jiushao obtained the equation⁸

$$x^{10} + 15x^8 + 72x^6 - 864x^4 - 11664x^2 - 34992 = 0.$$

The reasoning behind such a complicated equation is difficult to understand. Perhaps the approach to the problem was to equate two expressions for the area of the triangle formed by the center of the town, the tree, and the traveler. In that case, if the line from the traveler to the tree is represented as $b + \sqrt{a(x^2 + a)}$, the formula for the area of a triangle in terms of its sides is used, and the resulting area is equated to $\frac{1}{2}(a + (x^2)/2)b$, the result, after all radicals are cleared, will be an equation of degree 10 in x , but not the one mentioned by Qin Jiushao. It will be

$$ax^{10} + (a^2 - 4b^2)x^8 - 8abx^6 - 8a^2b^2x^4 + 16ab^4x^2 + 16a^2b^4 = 0.$$

One has to be very unlucky to get such a high-degree equation. Even a very simplistic approach leads only to a quartic equation. It is easy to see (Fig. 1) that if the diameter of the town rather than its square root is taken as the unknown, and the radius is drawn to the point of tangency, trigonometry will yield a quartic equation. If the radius is taken as the unknown, the similar right triangles in Fig. 1 lead to the cubic equation $2r^3 + 3r^2 = 243$. But, of course, the object of this game was probably to practice the art of algebra, not to get the simplest possible equation, no matter how virtuous it may seem to do so in other contexts. In any case, the historian's job is not that of a commentator trying to improve a text. It is to try to understand what the original author was thinking.

3. Japan

The Japanese mathematicians showed themselves to be superb algebraists from the beginning. We have already mentioned (Section 4 of Chapter 9) the quadrilateral problem of Sawaguchi Kazuyuki, which led to an equation of degree 1458, solved by Seki Kōwa. This problem, like many of the problems in the *sangaku* plaques,

⁸ Even mathematicians working within the Chinese tradition seem to have been puzzled by the needless elevation of the degree of the equation (Libbrecht, 1973, p. 136).

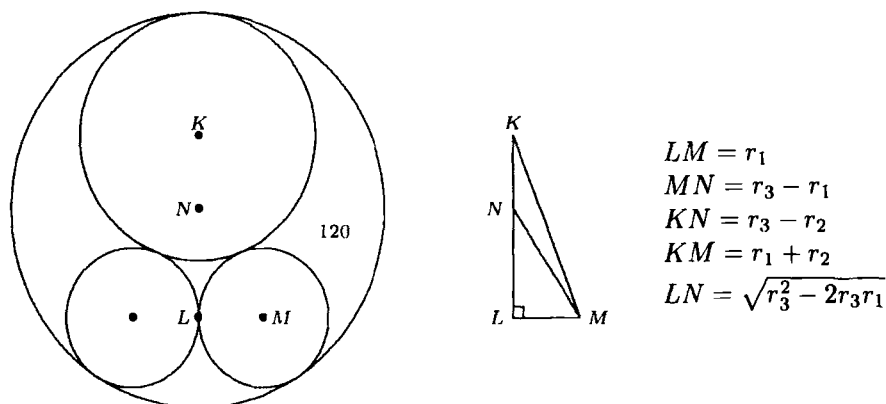


FIGURE 2. Sawaguchi Kazuyuki's first problem.

seems to be inspired by the desire to do some complicated algebra rather than by any pressing geometric need.

One impetus to the development of mathematics in Japan came with the arrival of the Chinese “method of the celestial element” (*tian yuan shu*), used in China. This name was given to the unknown in an equation by Li Ye in his 1248 treatise *Ceyuan Haijing* (*Sea Mirror of Circle Measurements*, see Mikami, 1913, p. 81).⁹ This term spread to Korea as *ch'onwonsul* and thence to Japan as *tengen jutsu*. This Chinese algebra became part of the standard Japanese curriculum before the seventeenth century.

Fifteen problems were published by Sawaguchi Kazuyuki in his 1670 work *Kokon Sampō-ki* (*Ancient and Modern Mathematics*). As an example of the great difficulty of these problems, consider the first of them. In this problem there are three circles each externally tangent to the other two and internally tangent to a fourth circle, as in Fig. 2. The diameters of two of the enclosed circles are equal and the third enclosed circle has a diameter five units larger. The area inside the enclosing circle and outside the three smaller circles is 120 square units. The problem is to compute the diameters of all four circles. This problem, although it yields to modern algebra, is complicated. In fact, Fig. 2 shows that the problem leads to the simultaneous equations

$$\begin{aligned} r_1 + \frac{5}{2} &= r_2, \\ 2\pi r_1^2 + \pi r_2^2 + 120 &= \pi r_3^2, \\ 4r_2^2 r_3 + 2r_1 r_2 r_3 + r_1 r_2^3 + r_1 r_3^2 &= 4r_2 r_3^2, \end{aligned}$$

where r_1 , r_2 , and r_3 are the radii of the circles. The last of these relations results from applying the Pythagorean theorem first to the triangle LMN to get LM , then to KLM .

3.1. Seki Kōwa. This problem was solved by Seki Kōwa (Smith and Mikami, 1914, pp. 96–97). In case Seki Kōwa's prowess in setting up and solving equations was not clear from his solution of Sawaguchi Kazuyuki's first problem, remember

⁹ The same word was used in a rather different and obscure sense by Qin Jiushao a year earlier in his *Sushu Jiu Zhang* (Libbrecht, 1973, pp. 345–346).

that he also solved the fourteenth problem (see Section 4 of Chapter 9), which involved an equation of degree 1458. Although the procedure was a mechanical one using counting boards, prodigious concentration must have been required to execute it. What a chess player Seki Kōwa could have been! As Mikami (1913, p. 160) remarks, "Perseverance and hard study were a part of the spirit that characterized Japanese mathematics of the old times."

Seki Kōwa was primarily an algebraist who converted the celestial element method into two sophisticated techniques for handling equations, known as *the method of explanation* and *the method of clarifying things of obscure origin*. He kept the latter method a secret. According to some scholars, his pupil Takebe Kenkō (1664–1739) refused to divulge the secret, saying, "I fear that one whose knowledge is so limited as mine would tend to misrepresent its significance." However, other scholars claim that Takebe Kenkō did write an exposition of the latter method, and that it amounts to the principles of cancellation and transposition. (See Section 2 of Chapter 3.)

Determinants. Seki Kōwa is given the credit for inventing one of the central ideas of modern mathematics: determinants. He introduced this subject in 1683 in *Kai Fukudai no Hō (Method of Solving Fukudai Problems)*.¹⁰ Nowadays determinants are usually introduced in connection with linear equations, but Seki Kōwa developed them in relation to equations of higher degree as well. The method is explained as follows. Suppose that we are trying to solve two simultaneous quadratic equations

$$\begin{aligned} ax^2 + bx + c &= 0 \\ a'x^2 + b'x + c' &= 0. \end{aligned}$$

When we eliminate x^2 , we find the linear equation

$$(a'b - ab')x + (a'c - ac') = 0.$$

Similarly, if we eliminate the constant term from the original equations and then divide by x , we find

$$(ac' - a'c)x + (bc' - b'c) = 0.$$

Thus from two quadratic equations we have derived two linear equations. Seki Kōwa called this process *tatamu (folding)*.

We have written out expressions for the simple 2×2 determinants here. For example,

$$\begin{vmatrix} a & c \\ a' & c' \end{vmatrix} = ac' - a'c;$$

but, as everyone knows, the full expanded expressions for determinants are very cumbersome even for the 3×3 case. It is therefore important to know ways of simplifying such determinants, using the structural properties we now call the *multilinear property* and the *alternating property*. Seki Kōwa knew how to make use of the multilinear property to take out a common factor from a given row. He not only formulated the concept of a determinant but also knew many of their properties, including how to determine which terms are positive and which are negative in the expansion of a determinant. It is interesting that determinants were introduced in

¹⁰ The word *fukudai* seems to be related to *fukugen suru*, meaning *reconstruct* or *restore*. According to Smith and Mikami (1914, p. 124), Seki Kōwa's school offered five levels of diploma, the third of which was called the *fukudai menkyo (fukudai license)* because it involved knowledge of determinants.

Europe around the same time (1693, by Leibniz), but in a comparatively limited context. As Smith and Mikami say (1914, p. 125),

It is evident that Seki was not only the discoverer but that he had a much broader idea than that of his great German contemporary.

4. Hindu algebra

The promising symbolic notation of the *Bakhshali Manuscript* was not adopted immediately throughout the world of Hindu mathematics. In particular, *Aryabhata I* tended to work in prose sentences. He considered the problem of finding two numbers given their product and their difference and gave the standard recipe for solving it.

4.1. Brahmagupta. The techniques involved with the *kuttaka* (pulverizer) belong to algebra, but since they are applied in number theory, we discussed them in that connection in Section 4 of Chapter 7. Brahmagupta also considered many problems that require finding the lengths of lines partitioning a polygon into triangles and quadrilaterals.

Brahmagupta's algebra is done entirely in words; for example (Colebrooke, 1817, p. 279), his recipe for the cube of a binomial is:

The cube of the last term is to be set down; and, at the first remove from it, thrice the square of the last multiplied by the preceding; then thrice the square of the preceding term taken into that last one; and finally the cube of the preceding term. The sum is the cube.

In short, $(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$. This rule is used for finding successive approximations to the cube root, just as in China and Japan. Similarly, in Section 4 (Colebrooke, 1817, p. 346), he tells how to solve a quadratic equation:

Take the absolute number from the side opposite to that from which the square and simple unknown are to be subtracted. To the absolute number multiplied by four times the [coefficient of the] square, add the square of the [coefficient of the] middle term; the square root of the same, less the [coefficient of the] middle term, being divided by twice the [coefficient of the] square is the [value of the] middle term.

Here the "middle term" is the unknown, and this statement is a very involved description of what we write as the quadratic formula:

$$x = \frac{\sqrt{4ac + b^2} - b}{2a} \quad \text{when} \quad ax^2 + bx = c.$$

Brahmagupta does not consider equations of degree higher than 2.

4.2. Bhaskara II. In the five centuries between Brahmagupta and Bhaskara II (who will henceforth be referred to simply as Bhaskara), the idea of using symbols for the unknown in an equation seems to have taken hold in Hindu mathematics. In Section 4 of his *Vija Ganita* (*Algebra*) Bhaskara reports that the initial syllables of the names for colors "have been selected by venerable teachers for names of values of unknown quantities, for the purpose of reckoning therewith" (Colebrooke, 1817, p. 139). He proceeds to give the rules for manipulating expressions involving such quantities; for example, the rule that we would write as $(-x - 1) + (2x - 8) = x - 9$ is written

$$\begin{array}{r} ya \dot{1} \quad ru \dot{1} \\ ya 2 \quad ru \dot{8} \\ \text{Sum } ya 1 \quad ru \dot{9}, \end{array}$$

where the dots indicate negative quantities. The syllable *ya* is the first syllable of the word for *black*, and *ru* is the first syllable of the word for *species*.¹¹

By the time of Bhaskara, the distinction between a rational and an irrational square root was well known. The Sanskrit word is *carani*, according to the commentator Krishna (Colebrooke, 1817, p. 145), who defines it as a number "the root of which is required but cannot be found without residue." Bhaskara gives rules such as $\sqrt{8} + \sqrt{2} = \sqrt{18}$ and $\sqrt{8} - \sqrt{2} = \sqrt{2}$.

Bhaskara's algebraic rules go beyond what is taught even today as standard algebra. He says that a nonzero number divided by zero gives an infinite quotient.

This fraction [3/0], of which the denominator is cipher, is termed an infinite quantity.

In this quantity consisting of that which has cipher for its divisor, there is no alteration, though many be inserted or extracted; as no change takes place in the infinite and immutable GOD, at the period of the destruction or creation of worlds, though numerous orders of beings are absorbed or put forth. [Colebrooke, 1817, pp. 137-138]

Both the *Vija Ganita* and the *Lilavati* contain problems on simple interest in which an unknown principal is to be found given the rate of simple interest and the amount to which it accrues after a given time. These equations are linear equations in one unknown.

The *Lilavati* contains a collection of problems in algebra, which are sometimes stated as though they were intended purely for amusement. For example, the rule for solving quadratic equations is applied in the *Vija Ganita* (Colebrooke, 1817, p. 212) to find the number of arrows x that Arjuna (hero of the *Bhagavad Gita*) had in his quiver, given that he shot them all, using $\frac{1}{2}x$ to deflect the arrows of his antagonist, $4\sqrt{x}$ to kill his antagonist's horse, six to kill the antagonist himself, three to demolish his antagonist's weapons and shield, and one to decapitate him. In other words, $x = \frac{1}{2}x + 4\sqrt{x} + 10$.

Bhaskara gives a criterion for a quadratic equation to have two (positive) roots. He also says that "if the solution cannot be found in this way, as in the case of cubic

¹¹ There is no evidence that Bhaskara knew of Diophantus; the fact that both describe a power of the unknown using a word whose meaning is approximated by the English word *species* is simply a coincidence.

or quartic equations, it must be found by the solver's own ingenuity" (Colebrooke, 1817, pp. 207-208). That ingenuity includes some work that would nowadays be regarded as highly inventive, not to say suspect; for example (Colebrooke, 1817, p. 214), how to solve the equation

$$\frac{(0(x + \frac{1}{2}x))^2 + 2(0(x + \frac{1}{2}x))}{0} = 15.$$

Bhaskara warns that multiplying by zero does not make the product zero, since further operations are to be performed. Then he simply cancels the zeros, saying that, since the multiplier and divisor are both zero, the expression is unaltered. The result is the equation we would write as $\frac{9}{4}x^2 + 3x = 15$. Bhaskara clears the denominator and writes the equivalent of $9x^2 + 12x = 60$. Even if the multiplication by zero is interpreted as multiplication by an expression that is tending to zero, as a modern mathematician would like to do, this cancellation is not allowed, since the first term in the numerator is a higher-order infinitesimal than the second. Bhaskara is handling 0 here as if it were 1. Granting that operation, he does correctly deduce, by completing the square (adding 4 to each side), that $x = 2$.

5. The Muslims

It has always been recognized that Europe received algebra from the Muslims; the very word *algebra* (*al-jabr*) is an Arabic word meaning *transposition* or *restoration*. Its origins in the Muslim world date from the ninth century, in the work of Muhammed ibn Musa al-Khwarizmi (780-850), as is well established.¹² What is less certain is how much of al-Khwarizmi's algebra was original with him and how much he learned from Hindu sources. According to Colebrooke (1817, pp. lxiv-lxxx), he was well versed in Sanskrit and translated a treatise on Hindu computation¹³ into Arabic at the request of Caliph al-Mamun, who ruled from 813 to 833. Colebrooke cites the Italian writer Pietro Cossali¹⁴ who presented the alternatives that al-Khwarizmi learned algebra either from the Greeks or the Hindus and opted for the Hindus. These alternatives are a false dichotomy. We need not conclude that al-Khwarizmi took everything from the Hindus or that he invented everything himself. It is very likely that he expounded some material that he read in Sanskrit and added his own ideas to it. Rosen (1831, p. x) explains the difference in the preface to his edition of al-Khwarizmi's algebra text, saying that "at least the method which he follows in expounding his rules, as well as in showing their application, differs considerably from that of the Hindu mathematical writers."

¹² Colebrooke (1817, p. lxxiii) noted that a manuscript of this work dated 1342 was in the Bodleian Library at Oxford. Obviously, this manuscript could not be checked out, and Colebrooke complained that the library's restrictions "preclude the study of any book which it contains, by a person not enured to the temperature of apartments unvisited by artificial warmth." If he worked in the library in 1816, his complaint would be understandable: Due to volcanic ash in the atmosphere, there was no summer that year. This manuscript is the source that Rosen (1831) translated and reproduced.

¹³ It is apparently this work that brought al-Khwarizmi's name into European languages in the form *algorism*, now *algorithm*. A Latin manuscript of this work in the Cambridge University Library, dating to the thirteenth century, has recently been translated into English (Crossley and Henry, 1990).

¹⁴ His dates are 1748-1813. He was Bishop of Parma and author of *Origine, trasporto in Italia, primi progressi in essa dell' algebra* (*The Origins of Algebra and Its Transmission to Italy and Early Progress There*), published in Parma in 1797.

Colebrooke also notes (p. lxxi) that Mohammed Abu'l-Wafa al-Buzjani (940–998) wrote a translation or commentary on the *Arithmetica* of Diophantus. This work, however, is now lost. Apart from these possible influences of Greek and Hindu algebra, whose effect is difficult to measure, it appears that the progress of algebra in the Islamic world was an indigenous growth. We shall trace that growth through several of its most prominent representatives, starting with the man recognized as its originator, Muhammed ibn Musa al-Khwarizmi.

5.1. Al-Khwarizmi. Besides the words *algebra* and *algorithm*, there is a common English word whose use is traceable to Arabic influence (although it is not an Arabic word), namely *root* in the sense of a square or cube root or a root of an equation. The Greek picture of the square root was the side of a square, and the word *side* (*pleura*) was used accordingly. The Muslim mathematicians apparently thought of the root as the part from which the equation was generated and used the word *jadhr* accordingly. According to al-Daffa (1973, p. 80), translations into Latin from Greek use the word *latus* while those from Arabic use *radix*. In English the word *side* lost out completely in the competition.

Al-Khwarizmi's numbers correspond to what we call positive real numbers. Theoretically, such a number could be defined by any convergent sequence of rational numbers, but in practice some rule is needed to generate the terms of the sequence. For that reason, it is more accurate to describe al-Khwarizmi's numbers as positive *algebraic numbers*, since all of his numbers are generated by equations with rational coefficients. The absence of negative numbers prevented al-Khwarizmi from writing all quadratic equations in the single form "squares plus roots plus numbers equal zero" ($ax^2 + bx + c = 0$). Instead, he had to consider three basic cases and three others, in which either the square or linear term is missing. He described the solution of "squares plus roots equal numbers" by the example of "a square plus 10 roots equal 39 dirhems." (A dirhem is a unit of money.) Al-Khwarizmi's solution of this problem is to draw a square of unspecified size (the side of the square is the desired unknown) to represent the square (Fig. 3). To add 10 roots, he then attaches to each side a rectangle of length equal to the side of the square and width $2\frac{1}{2}$ (since $4 \cdot 2\frac{1}{2} = 10$). The resulting cross-shaped figure has, by the condition of the problem, area equal to 39. He then fills in the four corners of the figure (literally "completing the square"). The total area of these four squares is $4 \cdot (2\frac{1}{2})^2 = 25$. Since $39 + 25 = 64$, the completed square has side 8. Since this square was obtained by adding rectangles of side $2\frac{1}{2}$ to each side of the original square, it follows that the original square had side 3.

This case is the one al-Khwarizmi considers first and is the simplest to understand. His figures for the other two cases of quadratic equations are more complicated, but all are based on Euclid's geometric illustration of the identity $((a+b)/2)^2 + ((a-b)/2)^2 = ab$ (Fig. 17 of Chapter 10).

Al-Khwarizmi did not consider any cubic equations. Roughly the first third of the book is devoted to various examples of pure mathematical problems leading to quadratic equations, causing the reader to be somewhat skeptical of his claim to be presenting the material needed in commerce and law. In fact, there are no genuine applications of quadratic equations in the book. But if quadratic equations have no practical applications (outside of technology, of course), there are occasions when a practical problem requires solving linear equations. Al-Khwarizmi found many

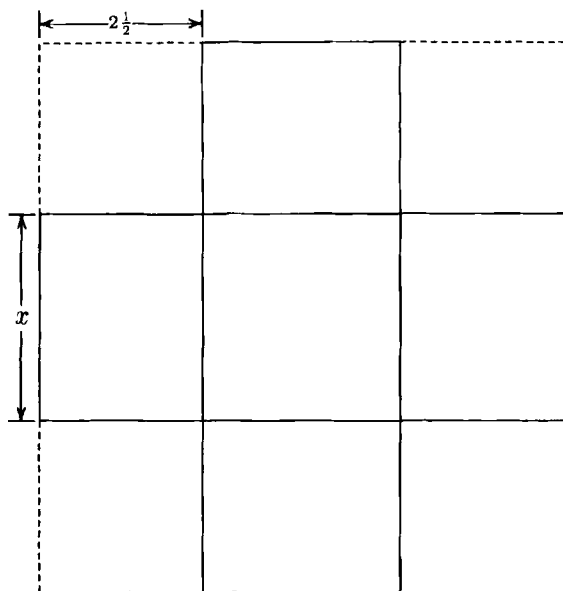


FIGURE 3. Al-Khwarizmi's solution of "square plus 10 roots equals 39 dirhems."

such cases in problems of inheritance, which occupy more than half of his *Algebra*. Here is a sample:

A man dies, leaving two sons behind him, and bequeathing one-fifth of his property and one dirhem to a friend. He leaves 10 dirhems in property and one of the sons owes him 10 dirhems. How much does each legatee receive?

Although mathematics is cross-cultural, its applications are very specific to the culture in which they are used. The difference between the modern solution of this legal problem and al-Khwarizmi's solution is considerable. Under modern law the man's estate would be considered to consist of 20 dirhems, the 10 dirhems cash on hand, and the 10 dirhems owed by one of the sons. The friend would be entitled to 5 dirhems (one-fifth plus one dirhem), and the indebted son would owe the estate 10 dirhems. His share of the estate would be one-half of the 15 dirhems left after the friend's share is taken out, or $7\frac{1}{2}$ dirhems. He would therefore have to pay $2\frac{1}{2}$ dirhems to the estate, providing it with cash on hand equal to $12\frac{1}{2}$ dirhems. His brother would receive $7\frac{1}{2}$ dirhems.

Now the notion of an estate as a legal entity that can owe and be owed money is a modern European one, alien to the world of al-Khwarizmi. Apparently in al-Khwarizmi's time, money could be owed only to a *person*. What principles are to be used for settling accounts in this case? Judging from the solution given by al-Khwarizmi, the estate is to consist of the 10 dirhems cash on hand, plus a *certain portion* (not all) of the debt the second son owed to his deceased father. This "certain portion" is the unknown in a linear equation and is the reason for invoking algebra in the solution. It is to be chosen so that *when the estate is divided up, the indebted son neither receives any more money nor owes any to the other heirs*. This condition leads to an equation that can be solved by algebra. Al-Khwarizmi

explains the solution as follows (we put the legal principle that provides the equation in capital letters):

Call the amount taken out of the debt *thing*. Add this to the property; the sum is 10 dirhems plus *thing*. Subtract one-fifth of this, since he has bequeathed one-fifth of his property to the friend. The remainder is 8 dirhems plus $\frac{4}{5}$ of *thing*. Then subtract the 1 dirhem extra that is bequeathed to the friend. There remain 7 dirhems and $\frac{4}{5}$ of *thing*. Divide this between the two sons. The portion of each of them is $3\frac{1}{2}$ dirhems plus $\frac{2}{5}$ of *thing*. THIS MUST BE EQUAL TO THING. Reduce it by subtracting $\frac{2}{5}$ of *thing* from *thing*. Then you have $\frac{3}{5}$ of *thing* equal to $3\frac{1}{2}$ dirhems. Form a complete *thing* by adding to this quantity $\frac{2}{3}$ of itself. Now $\frac{2}{3}$ of $3\frac{1}{2}$ dirhems is $2\frac{1}{3}$ dirhems, so that *thing* is $5\frac{5}{6}$ dirhems.

Rosen (1831, p. 133) suggested that the many arbitrary principles used in these problems were introduced by lawyers to protect the interests of next-of-kin against those of other legatees.

5.2. Abu Kamil. A commentary on al-Khwarizmi's *Algebra* was written by the mathematician Abu Kamil (ca. 850–930). His exposition of the subject contained none of the legacy problems found in al-Khwarizmi's treatise, but after giving the basic rules of algebra, it listed 69 problems of considerable intricacy to be solved. For example, a paraphrase of Problem 10 is as follows:

The number 50 is divided by a certain number. If the divisor is increased by 3, the quotient decreases by $3\frac{3}{4}$. What is the divisor?

Abu Kamil is also noteworthy because many of his problems were copied by Leonardo of Pisa, one of the first to introduce the mathematics of the Muslims into Europe.

5.3. Omar Khayyam. Although al-Khwarizmi did not consider any equations of degree higher than 2, such equations were soon to be considered by Muslim mathematicians. As we saw in Section 1 of Chapter 10, a link between geometry and algebra appeared in the use of the rectangular hyperbola by Pappus to carry out the *neûsis* construction for trisecting an angle (Fig. 9 of Chapter 10). The mathematician Omar Khayyam, of the late eleventh and early twelfth centuries (see Amir-Moez, 1963), realized that a large class of geometric problems of this type led to cubic equations that could be solved using conic sections. His treatise on algebra¹⁵ was largely occupied with the classification and solution of cubic equations by this method.

Omar Khayyam did not have modern algebraic symbolism. He lived within the confines of the universe constructed by the Greeks. His classification of equations, like al-Khwarizmi's, is conditioned by the use of only positive numbers as data. For that reason his classification is even more complicated than al-Khwarizmi's, since he is considering cubic equations as well as quadratics. He lists 25 types of equations (Kasir, 1931, pp. 51–52), six of which do not involve any cubic terms. The particular cubic we shall consider is *cubes plus squares plus sides equal number*, or, as

¹⁵ This treatise was little noticed in Europe until a French translation by Franz Woepcke (1827–1864) appeared in 1851 (Kasir, 1931, p. 7).

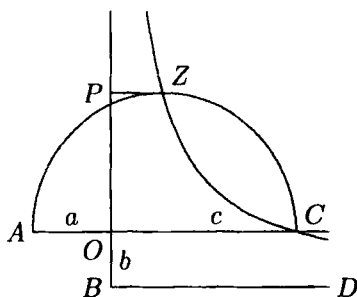


FIGURE 4. Omar Khayyam's solution of $x^3 + ax^2 + b^2x = b^2c$.

we would phrase it, $x^3 + ax^2 + bx = c$. In keeping with his geometric interpretation of magnitudes as line segments, Omar Khayyam had to regard the coefficient b as a square, so that we shall write b^2 rather than b . Similarly, he regarded the constant term as a solid, which without any loss of generality he considered to be a rectangular prism whose base was an area equal to the coefficient of the unknown. In keeping with this reduction we shall write b^2c instead of c . Thus Omar Khayyam actually considered the equation $x^3 + ax^2 + b^2x = b^2c$, where a , b , and c are data for the problem, to be represented as lines. His solution is illustrated in Fig. 4. He drew a pair of perpendicular lines intersecting at a point O and marked off $OA = a$ and $OC = c$ in opposite directions on one of the lines and $OB = b$ on the other line. He then drew a semicircle having AC as diameter, the line DB through B perpendicular to OB (parallel to AC), and the rectangular hyperbola through C having DB and the extension of OB as asymptotes. This hyperbola intersects the semicircle in the point C and in a second point Z . From Z he drew ZP perpendicular to the extension of OB , and ZP represented the solution of the cubic.

When it comes to actually producing a root by numerical procedures, Omar Khayyam's solution is circular, a mere restatement of the problem. He has broken the cubic equation into two quadratic equations in two unknowns, but any attempt to eliminate one of the two unknowns merely leads back to the original problem. In fact, no method of solution exists or can exist that reduces the solution of every cubic equation with real roots to the extraction of real square and cube roots of real numbers. What Omar Khayyam had created was an *analysis* of cubic equations using conic sections. He said that no matter how hard you look, you will never find a numerical solution "because whatever is obtained by conic sections cannot be obtained by arithmetic" (Amir-Moez, 1963, p. 336).

5.4. Sharaf al-Din al-Muzaffar al-Tusi. A generation after the death of Omar Khayyam, another Muslim mathematician, Sharaf al-Tusi (ca. 1135–1213, not to be confused with Nasir-Eddin al-Tusi, whose work was discussed in Section 4 of Chapter 11), wrote a treatise on equations in which he analyzed the cubic equation using methods that are surprisingly modern in appearance. This work has been analyzed by Hogendijk (1989). Omar Khayyam had distinguished between the eight types of cubic equations that always have a solution and the five that could fail to have a solution. Al-Tusi provided a numerical method of solution for the first eight types that was essentially the Chinese method of solving cubic equations.

He then turned to the five types that could have no (positive) solutions for some values of the data. As an example, one of these forms is

$$x^3 + ax^2 + c = bx.$$

For each of these cases, al-Tusi considered a particular value of x , which for this example is the value m satisfying

$$3m^2 + 2am = b.$$

Let us denote the positive root of this equation (the larger root, if there are two) by m . The reader will undoubtedly have noticed that the equation can be obtained by differentiating the original equation and setting x equal to m . The point m is thus in all cases a relative minimum of the difference of the left- and right-hand sides of the equation. That, of course, is precisely the property that al-Tusi wanted. Hogendijk comments that it is unlikely that al-Tusi had any concept of a derivative. In fact, the equation for m can be derived without calculus, by taking m as the value at which the minimum occurs, subtracting the values at x from the value at m , and dividing by $x - m$. The result is the inequality $m^2 + mx + x^2 + a(m + x) > b$ for $x > m$ and the opposite inequality for $x < m$. Therefore equality must hold when $x = m$, that is, $3m^2 + 2am = b$, which is the condition given by al-Tusi. After finding the point m , al-Tusi concluded that there will be no solutions if the left-hand side of the equation is larger than the right-hand side when $x = m$. There will be one unique solution, namely $x = m$ if equality holds there. That left only the case when the left-hand side was smaller than the right-hand side when $x = m$. For that case he considered the auxiliary cubic equation

$$y^3 + py^2 = d,$$

where p and d were determined by the type of equation. The quantity d was simply the difference between the right- and left-hand sides of the equation at $x = m$, that is, $bm - m^3 - am^2 - c$ in the present case, with p equal to $3m + a$. Al-Tusi was replacing x with $y = x - m$ here. The procedure was precisely the method we know as Horner's method, and the linear term drops out because the condition by which m was chosen ordains that it be so (see Problem 14.9.) The equation in y was known to have a root because it was one of the other 13 types, which always have solutions. Thus, it followed that the original equation must also have a solution, $x = m + y$, where y was the root of the new equation. The added bonus was that a lower bound of m was obtained for the solution.

6. Europe

As soon as translations from Arabic into Latin became generally available in the twelfth and thirteenth centuries, Western Europeans began to learn about algebra. The first of these was a Latin translation of al-Khwarizmi's *Algebra*, made in 1145 by Robert of Chester (dates unknown). Several talented mathematicians appeared early on who were able to make original contributions to its development. In some cases the books that they wrote were not destined to be published for many centuries, but at least one of them formed part of an Italian tradition of algebra that continued for several centuries.

6.1. Leonardo of Pisa (Fibonacci). Many of the problems in the *Liber abaci* reflect the routine computations that must be performed when converting currencies. These are applications of the Rule of Three that we have found in Brahmagupta and Bhaskara. Many of the other problems are purely fanciful. Leonardo's indebtedness to Arabic sources was detailed by Levey (1966, pp. 217–220), who listed 29 problems in the *Liber abaci* that are identical to problems in the *Algebra* of Abu Kamil. In particular, the problem of separating the number 10 into two parts satisfying an extra condition occurs many times. For example, one problem is to find x such that $10/x + 10/(10 - x) = 6\frac{1}{4}$.

The Liber quadratorum. The *Liber quadratorum* is written in the spirit of Diophantus. The resemblance in some points is so strong that it would be very strange if Leonardo had not seen a copy of Diophantus. This question is discussed by the translator of the *Liber quadratorum* (Sigler, 1987, pp. xi–xii), who notes that strong resemblances have been pointed out between the *Liber quadratorum* and al-Karaji's *Fakhri*, parts of which were copied from the *Arithmetica*, but that there are also parts of the *Liber quadratorum* that are original. The resemblance to Diophantus is shown in such statements as the ninth of its 24 propositions: *Given a nonsquare number that is the sum of two squares, find a second pair of squares having this number as their sum.* Leonardo's solution of this problem, like that of Diophantus, involves a great deal of arbitrariness, since the problem does not have a unique solution.

One advance in the *Liber quadratorum* is the use of general letters in an argument. Although in some proofs Leonardo argues much as Diophantus does, using specific numbers, he becomes more abstract in others. For example, Proposition 5 requires finding two numbers the sum of whose squares is a square that is also the sum of the squares of two given numbers (Problem 9 of Book 2 of Diophantus). He says to proceed as follows. Let the two given numbers be $.a.$ and $.b.$ and the sum of their squares $.g.$. Now take any other two numbers $.de.$ and $.ez.$ [not proportional to the given numbers] the sum of whose squares is a square. These two numbers are arranged as the legs of a right triangle. If the square on the hypotenuse of this triangle is $.g.$, the problem is solved. If the square on the hypotenuse is larger than $.g.$, mark off the square root of $.g.$ on the hypotenuse. The projections (as we would call them) of this portion of the hypotenuse on each of the legs are known, since their ratios to the square root of $.g.$ are known. Moreover, that ratio is rational, since they are the same as the ratios of $.a.$ and $.b.$ to the hypotenuse of the original triangle. These two projections therefore provide the new pair of numbers. Being proportional to $.a.$ and $.b.$, which are not proportional to the two numbers given originally, they must be different from those numbers. This argument is more convincing, because more abstract, than proofs by example, but the geometric picture plays an important role in making the proof comprehensible.

The Flos. Leonardo's approach to algebra begins to look modern in other ways as well. In one of his works, called the *Flos super solutionibus quarundam questionum ad numerum et ad geometriam vel ad utrumque pertinentum* (*The Full Development*¹⁶ of the Solutions of Certain Questions Pertaining to Number or Geometry or Both, see Boncompagni, 1854, p. 4), he mentions a challenge from John of Palermo to find a number satisfying $x^3 + 2x^2 + 10x = 20$ using the methods

¹⁶ The word *flos* means *bloom*, and is used in the figurative sense of "the bloom of youth." That appears to be its meaning here.

given by Euclid in Book 10 of the *Elements*, that is, to construct a line of this length using ruler and compass. In working on this question, Leonardo made two important contributions to algebra, one numerical and one theoretical. The numerical contribution was to give the unique positive root in sexagesimal notation correct to six places. The theoretical contribution was to show by using divisibility properties of numbers that there cannot be a rational solution or a solution obtained using only rational numbers and square roots of rational numbers.

6.2. Jordanus Nemorarius. The translator and editor of the book *De numeris datis* (*On Given Numbers*), written by Jordanus Nemorarius, says (Hughes, 1981, p. 11) "It is reasonable to assume... that Jordanus was influenced by al-Khwarizmi's work." This conclusion was reached on the basis of Jordanus' classification of quadratic equations and his order of expounding the three types, among other resemblances between the two works.

De numeris datis is the algebraic equivalent of Euclid's *Data*. Where Euclid says that a line is given (determined) if its ratio to a given line is given, Jordanus Nemorarius says that a number is given if its ratio to a given number is given. The well-known elementary fact that two numbers can be found if their sum and difference are known is generalized to the theorem that any set of numbers can be found if the differences of the successive numbers and the sum of all the numbers is known.¹⁷ In general, this book contains a large variety of data sets that determine numbers. For example, *if the sum of the squares of two numbers is known, and the square of the difference of the numbers is known, the numbers can be found*. The four books of *De numeris datis* contain about 100 such results. These results admit a purely algebraic interpretation. For example, in Book 4 Jordanus Nemorarius writes:

If a square with the addition of its root multiplied by a given number makes a given number, then the square itself will be given.
[Hughes, 1981, p. 100]¹⁸

Where earlier mathematicians would have proved this proposition with examples, Jordanus Nemorarius uses letters representing abstract numbers. The assertion is that there is only one (positive) number x such that $x^2 + \alpha x = \beta$, and that x can be found if α and β are given.

6.3. The fourteenth and fifteenth centuries. The century in which Nicole d'Oresme made such remarkable advances in geometry, coming close to the creation of analytic geometry, was also a time of rapid advance in algebra, epitomized by Antonio de' Mazzinghi (ca. 1353-1383). His *Trattato d'algebra* contains some complicated systems of linear and quadratic equations in as many as three unknowns (Franci, 1988). He was one of the earliest algebraists to move the subject toward the numerical and away from the geometric interpretation of problems.

In the following century Luca Pacioli wrote *Summa de arithmetica, geometrica, proportioni et proportionalita* (*Treatise on Arithmetic, Geometry, Proportion, and Proportionality*), which was closer to the elementary work of al-Khwarizmi and more geometrical in its approach to algebra than the work of Mazzinghi. Actually

¹⁷ This statement is a variant of the *epanthēma* (blossom) of Thymaridas.

¹⁸ This translation is my own and is intended to be literal; Hughes gives a smoother, more idiomatic translation on p. 168.

(see Parshall, 1988), the work was largely a compilation of the works of Leonardo of Pisa, but it did bring the art of abbreviation closer to true symbolic notation. For example, what we now write as $x - \sqrt{x^2 - 36}$ was written by Pacioli as

$$1.co.\tilde{m}Rv.1.ce \tilde{m}36.$$

Here *co* means *cosa* (*thing*), the unknown; *ce* means *censo* (*power*), and *Rv* is probably a printed version of *Rx*, from the Latin *radix*, meaning *root*.¹⁹ Pacioli's work was both an indication of how widespread knowledge of algebra had become by this time and an important element in propagating it. The sixteenth-century Italian algebraists who moved to the forefront of the subject and advanced it far beyond where it had been up to that time had all read Pacioli's treatise thoroughly.

6.4. Chuquet. The *Triparty en la science des nombres* by Nicolas Chuquet is accompanied by a book of problems to illustrate its principles, a book on geometrical mensuration, and a book of commercial arithmetic. The last two are applications of the principles in the first book. Thus the subject matter is similar to that of al-Khwarizmi's *Algebra* or Leonardo's *Liber abaci*.

There are several new things in the *Triparty*. One is a superscript notation similar to the modern notation for the powers of the unknown in an equation. The unknown itself is called the *premier* or "first." Algebra in general is called the *rigle des premiers* or "rule of firsts." Chuquet listed the first 20 powers of 2 and pointed out that when two such numbers are multiplied, their indices are added. Thus, he had a clear idea of the laws of integer exponents. A second innovation in the *Triparty* is the free use of negative numbers as coefficients, solutions, and exponents. Still another innovation is the use of some symbolic abbreviations. For example, the square root is denoted R^2 (*R* for the Latin *radix*, or perhaps the French *racine*). The equation we would write as $3x^2 + 12 = 9x$ was written $.3.^2 \tilde{p}.12. \text{ egaulx a } .9.^1$. Chuquet called this equation impossible, since its solution would involve taking the square root of -63 .

His instructions are given in words. For example (Struik, 1986, p. 62), consider the equation

$$R^2 4^2 \tilde{p}.4^1 \tilde{p}.2^1 \tilde{p}.1 \text{ egaulx a } .100,$$

which we would write

$$\sqrt{4x^2 + 4x + 2x + 1} = 100.$$

Chuquet says to subtract $.2^1 \tilde{p}.1$ from both sides, so that the equation becomes

$$R^2 4^2 \tilde{p}.4^1 \text{ egaulx a } .99\tilde{m}.2^1.$$

Next he says to square, getting

$$4^2 \tilde{p}.4^1 \text{ egaulx a } 9801.\tilde{m}.396^1 \tilde{p}.4^2.$$

Subtracting 4^2 from both sides and adding 396.1 to both sides then yields

$$400^1 \text{ egaulx a } .9801..$$

Thus $x = 9801/400$.

¹⁹ The symbol *Rx* should not be confused with the same symbol in pharmacy, which comes from the Latin *recipe*, meaning *take*.

6.5. Solution of cubic and quartic equations. In Europe algebra was confined to linear and quadratic equations for many centuries, whereas the Chinese and Japanese had not hesitated to attack equations of any degree. The difference in the two approaches is a result of different ideas of what constitutes a solution. This distinction is easy to make nowadays: The European mathematicians were seeking an exact solution using only arithmetic operations and root extractions, what is called *solution by radicals*. However, it will not do to press the distinction too far: It is impossible to do good numerical work without a sound theoretical basis. As we saw in the work of Sharaf al-Tusi, the coefficients that appear in the course of his numerical solution have theoretical significance.

The Italian algebraists of the early sixteenth century made advances in the search for a general algorithm for solving higher-degree equations. We discussed the interesting personal aspects of the solution of cubic equations in Section 4 of Chapter 3. Here we concentrate on the technical aspects of the solution.

The verses Tartaglia had memorized say, in modern language, that to solve the problem $x^3 + px = q$, one should look for two numbers u and v satisfying $u - v = q$, $uv = (p/3)^3$. The problem of finding u and v is that of finding two numbers given their difference and their product, and of course, that is merely a matter of solving a *quadratic* equation, a problem that had already been completely solved. Once this quadratic has been solved, the solution of the original cubic is $x = \sqrt[3]{u} - \sqrt[3]{v}$. The solution of the cubic has thus been reduced to solving a quadratic equation, taking the cube roots of its two roots, and subtracting. Cardano illustrated with the case of "a cube and six times the side equal to 20." Using his complicated rule (complicated because he stated it in words), he gave the solution as

$$\sqrt[3]{\sqrt{108} + 10} - \sqrt[3]{\sqrt{108} - 10}.$$

He did not add that this number equals 2.

Ludovico Ferrari. Cardano's student Ludovico Ferrari worked with him in the solution of the cubic, and between them they had soon found a way of solving certain fourth-degree equations. Ferrari's solution of the quartic was included near the end of Cardano's *Ars magna*. Counting cases as for the cubic, one finds a total of 20 possibilities. The principle in most cases is the same, however. The idea is to make a perfect square in x^2 equal to a perfect square in x by adding the same expression to both sides. Cardano gives the example

$$60x = x^4 + 6x^2 + 36.$$

It is necessary to add to both sides an expression $rx^2 + s$ to make them squares, that is, so that both sides of

$$rx^2 + 60x + s = x^4 + (6 + r)x^2 + (36 + s)$$

are perfect squares. Now the condition for this to happen is well known: $ax^2 + bx + c$ is a perfect square if and only if $b^2 - 4ac = 0$. Hence we need to have simultaneously

$$3600 - 4sr = 0, \quad (6 + r)^2 - 4(36 + s) = 0.$$

Solving the second of these equations for s in terms of r and substituting in the first leads to the equation

$$r^3 + 12r^2 = 108r + 3600.$$

This is a cubic equation called the *resolvent* cubic. Once it is solved, the original quartic breaks into two quadratic equations upon taking square roots and adding an ambiguous sign.

A few aspects of the solution of cubic and quartic equations should be noted. First, the problem is not a practical one. Second, the Cardano recipe for solving an equation sometimes gives the solution in a rather strange form. For example, Cardano says that the solution of $x^3 + 6x = 20$ is $\sqrt[3]{\sqrt{108} + 10} - \sqrt[3]{\sqrt{108} - 10}$. The expression is correct, but can you tell at a glance that it represents the number 2?

Third, the procedure does not always work. For example, the equation $x^3 + 6 = 7x$ has to be solved by guessing a number that can be added to both sides so as to produce a common factor that can be canceled out. The number in this case is 21, but there is no *algorithm* for finding such a number. For equations of this type the algebraic procedures for finding x involve square roots of negative numbers. The search for an algebraic procedure using only real numbers to solve this case of the cubic continued for 300 years, until finally it was shown that no such procedure can exist.

6.6. Consolidation. There were two natural ways to build on what had been achieved in algebra by the end of the sixteenth century. One was to find a notation that could unify equations so that it would not be necessary to consider so many different cases and so many different possible numbers of roots. The other was to solve equations of degree five and higher. We shall discuss the first of these here and devote Chapter 15 to the quest for the second and its consequences.

All original algebra treatises written up to and including the treatise of Bombelli are very tiresome for the modern student, who is familiar with symbolic notation. For that reason we have sometimes allowed ourselves the convenience of modern notation when doing so will not distort the thought process involved. In the years between 1575 and 1650 several innovations in notation were introduced that make treatises written since that time appear essentially modern. The symbols $+$ and $-$ were originally used in bookkeeping in warehouses to indicate excess and deficiencies; they first appeared in a German treatise on commercial arithmetic in 1489 but were not widely used in the rest of Europe for another century. The sign for equality was introduced by a Welsh medical doctor, physician to the short-lived Edward VI, named Robert Recorde (1510–1558). His symbol was a very long pair of parallel lines, because, as he said, “noe 2. thynges, can be moare equalle.” The use of abbreviations for the various powers of the unknown in an equation was eventually achieved, but there were two other needs to be met before algebra could become a mathematical subject on a par with geometry: a unified way of writing equations and a concept of number in which every equation would have a solution. The use of exponential notation and grouping according to powers was discussed by Simon Stevin (see Section 7 of Chapter 6). Stevin used the abbreviation M for the first unknown in a problem, sec for the second, and ter for the third. Thus (see Zeuthen, 1903, p. 95), what we would write as the equation

$$\frac{6x^3}{y} \div 2xz^2 = \frac{3x^2}{yz^2}$$

was expressed as follows: If we divide

$$6 M \text{ ③ } D sec \text{ ① } \text{ by } 2 M \text{ ① } ter \text{ ② } ,$$

we obtain

$$3 M \textcircled{2} D \textit{sec} \textcircled{1} D \textit{ter} \textcircled{2} .$$

Although notation still had far to go, from the modern point of view, at least it was no longer necessary to use a different letter to represent each power of the unknown in a problem.

François Viète. The French lawyer François Viète (1540–1603), who worked as tutor in a wealthy family and later became an advisor to Henri de Navarre (the future king Henri IV), found time to study Diophantus and to introduce his own ideas into algebra. Viète is credited with several crucial advances in the subject. In his book *Artis analyticae praxis* (*The Practice of the Analytic Art*) he begins by giving the rules for powers of binomials (in words). For example, he describes the fifth power of a binomial as “the fifth power of the first [term], plus the product of the fourth power of the first and five times the second, . . .” Viète’s notation was slightly different from ours, but is more recognizable to us than that of Stevin. He would write the equation $A^3 + 3BA = D$, where the vowel A represented the unknown and the consonants B and D were taken as known, as follows (Zeuthen, 1903, p. 98):

$$A \textit{cubus} + B \textit{planum in} A3 \textit{aequatur} D \textit{solido}.$$

As this quotation shows, Viète appears to be following the tedious route of writing everything out in words, and to be adhering to the requirement that all the terms in an equation be geometrically homogeneous.

This introduction is followed by five books of *zetetics* (research, from the Greek word *zētein*, meaning *seek*). The mention of “roots” in connection with the binomial expansions was not accidental. Viète studied the relation between roots and coefficients in general equations. By using vowels to represent unknowns and consonants to represent data for a problem, Viète finally achieved what was lacking in earlier treatises: a convenient way of talking about general data without having to give specific examples. His consonants could be thought of as representing numbers that would be known in any particular application of a process, but were left unspecified for purposes of describing the process itself. His first example was the equation $A^2 + AB = Z^2$, in other words, a standard quadratic equation. According to Viète these three letters are associated with three numbers in direct proportion, Z being the middle, B the difference between the extremes, and A the smallest number. In our terms, $Z = Ar$ and $B = Ar^2 - A$. Thus, the general problem reduces to finding the smallest of three numbers A , Ar , Ar^2 given the middle value and the difference of the largest and smallest. Viète had already shown how to do that in his books of *zetetics*.

This analysis showed Viète the true relation between the coefficients and the roots. For example, he knew that in the equation $x^3 - 6x^2 + 11x = 6$, the sum and product of the roots must be 6 and the sum of the products taken two at a time must be 11. This observation still did not enable him to solve the general cubic equation, but he did study the problem geometrically and show that any cubic could be solved provided that one could solve two of the classical problems of antiquity: constructing two mean proportionals between two given lines and trisecting any angle. As he concluded at the end of his geometric chapter: “It is very worthwhile to note this.”

Questions and problems

14.1. Problem 6 of Book 1 of the *Arithmetica* is to separate a given number into two numbers such that a given fraction of the first exceeds a given fraction of the other by a given number. In our terms this is a problem in two unknowns x and y , and there are four bits of data: the sum of the two numbers, which we denote by a , the two proper fractions r and s , and the amount b by which rx exceeds sy . Write down and solve the two equations that this problem involves. Under what conditions will the solutions be positive rational numbers (assuming that a , b , r , and s are positive rational numbers)? Compare your statement of this condition with Diophantus' condition, stated in very complicated language: *The last given number must be less than that which arises when that fraction of the first number is taken which exceeds the other fraction.*

14.2. Carry out the solution of the bundles of wheat problem from the *Jiu Zhang Suanshu*. Is it possible to solve this problem without the use of negative numbers?

14.3. Solve the equation for the diameter of a town considered by Li Rui. [Hint: Since $x = -3$ is an obvious solution, this equation can actually be written as $x^3 + 3x^2 = 972$.]

14.4. Solve the following legacy problem from al-Khwarizmi's *Algebra*: *A woman dies and leaves her daughter, her mother, and her husband, and bequeaths to some person as much as the share of her mother and to another as much as one-ninth of her entire capital. Find the share of each person.* It was understood from legal principles that the mother's share would be $\frac{2}{13}$ and the husband's $\frac{3}{13}$.

14.5. Solve the problem of Abu Kamil in the text.

14.6. If you know some modern algebra, explain, by filling in the details of the following argument, why it is not surprising that Omar Khayyam's geometric solution of the cubic cannot be turned into an algebraic procedure. Consider a cubic equation with rational coefficients but no rational roots,²⁰ such as $x^3 + x^2 + x = 2$. By Omar Khayyam's method, this equation is replaced with the system $y(z + 1) = 2$, $z^2 = (y + 1)(2 - y)$, one obvious solution of which is $y = 2$, $z = 0$. The desired value of x is the y -coordinate of the other solution. The procedure for eliminating one variable between the two quadratic equations representing the hyperbola and circle is a rational one, involving only multiplication and addition. Since the coefficients of the two equations are rational, the result of the elimination will be a polynomial equation with rational coefficients. If the root is irrational, that polynomial will be divisible by the minimal polynomial for the root over the rational numbers. However, a cubic polynomial with rational coefficients but no rational roots is itself the minimal polynomial for all of its roots. Hence the elimination will only return the original problem.

14.7. Why did al-Khwarizmi include a complete discussion of the solution of quadratic equations in his treatise when he had no applications for them at all?

14.8. Contrast the modern Western solution of the Islamic legacy problem discussed in the text with the solution of al-Khwarizmi. Is one solution "fairer" than the other? Can mathematics make any contribution to deciding what is fair?

²⁰ If the coefficients are rational, their denominators can be cleared. Then all rational roots will be found among the finite set of fractions whose numerators divide the constant term and whose denominators divide the leading coefficient. There is an obvious algorithm for finding these roots.

14.9. Consider the cubic equation of Sharaf al-Tusi's third type, which we write as $-x^3 - ax^2 + bx - c = 0$. Using Horner's method, as described in Section 2, show that if the first approximation is $x = m$, where m satisfies $3m^2 + 2am - b = 0$, then the equation to be satisfied at the second approximation is $y^2 - (3m + a)y^2 - (m^3 + am^2 - bm + c) = 0$. That is, carry out the algorithm for reduction and show that the process is

$$\begin{array}{rcl} & -c & -m^3 - am^2 + bm - c \\ b & \longrightarrow & -3m^2 - 2am + b (= 0) \\ -a & & -3m - a \\ -1 & & -1 \end{array}$$

14.10. Consider Problem 27 of Book 1 of *De numeris datis*: Two numbers are given whose sum is 10. If one is divided by 4 and the other by 2, the product of the quotients is 2. What are the two numbers? Solve this problem in your own way, then solve it following Jordanus' recipe, which we paraphrase as follows. Let the two numbers be x and y , and let the quotients be e and f when x and y are divided by c and d respectively; let the product of the quotients be $ef = b$. Let $bc = h$, which is the same as fce or fx . Then multiply d by h to produce j , which is the same as xd or xy . Since we now know both $x + y$ and xy , we can find x and y .

14.11. Solve the equation $x^3 + 60x = 992$ using the recipe given by Tartaglia.

14.12. How can you prove that $\sqrt[3]{\sqrt{108} + 10} - \sqrt[3]{\sqrt{108} - 10} = 2$?

14.13. If you know the polar form of complex numbers $z = r \cos \theta + ir \sin \theta$, show that the problem of taking the cube root of a complex number is equivalent to solving two of the classical problems of antiquity simultaneously, just as Viète claimed: the problem of two mean proportionals and the problem of trisecting the angle.

14.14. Consider Viète's problem of finding three numbers in direct proportion given the middle number and the difference between the largest and smallest. Show that this problem amounts to finding x and y given \sqrt{xy} and $y - x$. How do you solve such a problem?

14.15. Show that the equation $x^3 = px + q$, where $p > 0$ and $q > 0$, has the solution $x = \sqrt{4p/3} \cos \theta$, where $\theta = \frac{1}{3} \arccos((q\sqrt{27})/(2\sqrt{p^3}))$. In order for this inverse cosine to exist it is necessary and sufficient that $q^2/4 - p^3/27 \leq 0$, which is precisely the condition under which the Cardano formula requires the cube root of a complex number. [Hint: Use the formula $4 \cos^3 \theta - 3 \cos \theta = \cos(3\theta)$.]

Observe that

$$\theta = \frac{1}{3} \int_a^1 \frac{1}{\sqrt{1-t^2}} dt,$$

where $a = (q\sqrt{27})/(2\sqrt{p^3})$. Thus, the solution of the cubic equation has a connection with the integral of an algebraic function $1/y$, where y satisfies the quadratic equation $y^2 = 1 - x^2$. This kind of connection turned out to be the key to the solution of higher-degree algebraic equations. As remarked in the text, Viète's solution of the cubic uses a transcendental method, even though an algebraic method exists.

CHAPTER 15

Modern Algebra

By the mid-seventeenth century, the relation between the coefficients and roots of a general equation was understood, and it was conjectured that if you counted roots according to multiplicity and allowed complex roots, an equation of degree n would have n roots. Algebra had been consolidated to the point that the main unsolved problem, the solution of equations of degree higher than 4, could be stated simply and analyzed.

The solution of this problem took nearly two centuries, and it was not until the late eighteenth and early nineteenth centuries that enough insight was gained into the process of determining the roots of an equation from its coefficients to prove that arithmetic operations and root extractions were not sufficient for this purpose. Although the solution was a negative result, it led to the important concepts of modern algebra that we know as groups, rings, and fields; and these, especially groups, turned out to be applicable in many areas not directly connected with algebra. Also on the positive side, nonalgebraic methods of solving higher-degree equations were also sought and found, and a theoretically perfect way of deciding whether a given equation can be solved in radicals was produced.

1. Theory of equations

Viète understood something of the relation between the roots and the coefficients of some equations. His understanding was not complete, because he was not able to find all the roots. Before the connection could be made completely, there had to be a domain in which an equation of degree n would have n roots. Then the general connection could be made for quadratic, cubic, and quartic equations and generalized from there. The missing theorem was eventually to be called the *fundamental theorem of algebra*.¹

1.1. Albert Girard. This fundamental theorem was first stated by Albert Girard (1595–1632), the editor of the works of Simon Stevin. In 1629 he wrote *L'invention nouvelle en l'algèbre* (*New Discovery (Invention) in Algebra*). This work contained some of the unifying concepts that make modern algebra the compact, efficient system that it now is. One of these, for example, is regarding the constant term as the coefficient of the zeroth power of the unknown. He introduced the notion of *factions* of a finite set of numbers. The first faction is the sum of the numbers, the second the sum of all products of two distinct numbers from the set, and so on. The last faction is the product of all the numbers, so that “there are as many

¹ In his textbook on analytic function theory (*Analytic Function Theory*, Ginn & Co., Boston, 1960, Vol. 1, p. 24), Einar Hille (1894–1980) wrote that “modern algebraists are inclined to deny both its algebraic and its fundamental character.” Hille does not name the modern algebraists, but he was a careful writer who must have had someone in mind. In the context of its time, the theorem was both algebraic and fundamental.

factions as there are numbers given." He noted that the number of terms in each faction could be found by using Pascal's triangle.

Girard always regarded the leading coefficient as 1. Putting the equation into this form, he stated as a theorem (see, for example, Struik, 1986, p. 85) that "all equations of algebra receive as many solutions as the denomination [degree] of the highest form shows, except the incomplete, and the first faction of the solutions is equal to the number of the first mixed [that is, the coefficient of the power one less than the degree of the equation], their second faction is equal to the number of the second mixed, their third to the third mixed, and so on, so that the last faction is equal to the closure [product], and this according to the signs that can be observed in the alternate order." This recognition that the coefficients of a polynomial are elementary symmetric polynomials in its zeros was the first ray of light at the dawn of modern algebra.

By "incomplete," Girard meant equations with some terms missing. In some cases, he said, these may not have a full set of solutions. He gave the example of the equation $x^4 = 4x - 3$, whose solutions he gave as 1, 1, $-1 + \sqrt{-2}$, and $-1 - \sqrt{-2}$, showing that he realized the need to count both complex roots and multiple real roots for the sake of the general rule. He invoked the simplicity of the general rule as justification for introducing the multiple and complex roots, along with the fact that complex numbers provide solutions where otherwise none would exist.

1.2. Tschirnhaus transformations. Every complex number has n th roots—exactly n of them except in the case of 0—that are also complex numbers. As a consequence, any formula for solving equations that involves only the application of rational operations and root extractions starting with the coefficients will remain within the domain of complex numbers. This elementary fact led to the proposition stated by Girard, which we know as the fundamental theorem of algebra. Finding such a formula for equations of degree five and higher was to become a preoccupation of algebraists for the next two centuries.

Analysis of the cubic. By the year 1600 equations of degrees 2, 3, and 4 could all be solved, assuming that one could extract the cube root of a complex number. The methods used suggest an inductive process in which the solution of an equation of degree n , say

$$x^n - a_1x^{n-1} + \cdots \mp a_{n-1}x \pm a_n = 0,$$

would be found by a substitution $y = x^{n-1} - b_1x^{n-2} + \cdots \pm b_{n-2}x \mp b_{n-1}$ with the coefficients b_1, \dots, b_{n-1} chosen so that the original equation becomes $y^n = C$. Observe that there are $n - 1$ coefficients b_k at our disposal and $n - 1$ coefficients a_1, \dots, a_{n-1} to be removed from the original equation. The program looks feasible. Something of the kind must have been the reasoning that led Ehrenfried Walther von Tschirnhaus (1652–1708) to the belief that he had discovered a general solution to all polynomial equations. In 1677 he wrote to Leibniz:

In Paris I received some letters from Mr. Oldenburg, but from lack of time have not yet been able to write back that I have found a new way of determining the irrational roots of all equations. . . The entire problem reduces to the following: We must be able to remove all the middle terms from any equation. When that is done, and as a result only a single power and a single known quantity remain, one need only extract the root.

Tschirnhaus claimed that the the middle terms (the a_k above) would be eliminated by a polynomial of the sort just discussed, provided that the b_k are suitably chosen. Such a change of variable is now called a *Tschirnhaus transformation*. If a Tschirnhaus transformation could be found for the *general* equation of degree n , and a formula existed for solving the *general* equation of degree $n - 1$, the two could be combined to generate a formula for solving the general equation of degree n . At the time, there was not even a Tschirnhaus transformation for the cubic equation. Tschirnhaus was to provide one.

He illustrated his transformation using the example $x^3 - qx - r = 0$. Taking $y = x^2 - ax - b$, he noted that y satisfied the equation

$$y^3 + (3b - 2q)y^2 + (3b^2 + 3ar - 4qb + q^2 - a^2q)y + (b^2 - 2qb^2 + 3bar + q^2b - aqr - a^2qb + a^3r - r^2) = 0.$$

He eliminated the square term by choosing $b = 2q/3$, then removed the linear term by solving for a in the quadratic equation

$$qa^2 - 3ra + 4q^2/3 = 0.$$

In this way, he had found at the very least a second solution of the general cubic equation, independent of the solution given by Cardano. And, what is more important, he had indicated a plausible way by which any equation might be solved. If it worked, it would prove that every polynomial equation could be solved using rational operations and root extractions, thereby proving at the same time that the complex numbers are algebraically closed. Unfortunately, detailed examination of the problem revealed difficulties that Tschirnhaus had apparently not noticed at the time of his letter to Leibniz.

The main difficulty is that when the variable x is eliminated between two polynomial equations $p_n(x) = 0$ and $y = p_{n-1}(x)$, where p_n is of degree n and p_{n-1} of degree $n - 1$, the degrees of the equations needed to eliminate the successive coefficients in the equation for y increase to $(n - 1)!$, not $n - 1$.² It is only in the case of a cubic, where $(n - 1)! = n - 1$, that the program can be made to work in general. It may, however, work for a *particular* equation of higher degree. Leibniz, at any rate, was not convinced. He wrote to Tschirnhaus,

I do not believe that [your method] will be successful for equations of higher degree, except in special cases. I believe that I have a proof for this. [Kracht and Kreyszig, 1990, p. 27]

Tschirnhaus' method had intuitive plausibility: If there existed an algorithm for solving all equations, that algorithm should be a procedure like the Tschirnhaus transformation. Because the method does *not* work, the thought suggests itself that there may be equations that cannot be solved algebraically. The work of Tschirnhaus and Girard had produced two important insights into the general problem of polynomial equations: (1) the coefficients are symmetric functions of the roots; (2) solving the equation should be a matter of finding a sequence of operations that would eliminate coefficients until a pure equation $y^n = C$ was obtained. Since the problem was still unresolved, still more new insights were needed.

² Seki Kōwa knew the rational procedures (what he called *folding*, as discussed in Section 3 of Chapter 14) for eliminating x . It does seem a pity that the contemporaries Tschirnhaus and Seki Kōwa lived so far apart. They would have had much to talk about if they could have met.

To explain the most important of these new insights, let us consider what Girard's result means when applied to Cardano's solution of the cubic $y^3 + py = q$. If the roots of this equation are r , s , and t , then $p = st + tr + rs$, $q = rst$, $t = -r - s$, since the coefficient of y^2 is zero. The sequence of operations implied by Cardano's formula is

$$\begin{aligned} u &= \frac{p}{3}; & v &= \frac{q}{2}; \\ a &= \sqrt{u^3 + v^2}; \\ y &= \sqrt[3]{v + a} + \sqrt[3]{v - a}. \end{aligned}$$

Girard's work implies that the quantity a , which is an *irrational* function of the coefficients p and q , is a *rational* function of the roots r , s , and t :

$$a = \pm \frac{i}{\sqrt{108}}(r - s)(s - t)(t - r);$$

that is, it does not involve taking the square root of any expression containing a root.

1.3. Newton, Leibniz, and the Bernoullis. In the 1670s Newton wrote a text-book of algebra called *Arithmetica universalis*, which was published in 1707, in which he stated more clearly and generally than Girard had done the relation between the coefficients and roots of a polynomial. Moreover, he showed that other symmetric polynomials of the roots could be expressed as polynomials in the coefficients by giving a set of rules that are still known by his name, although Edward Waring also proved that such an expression is possible.

Another impetus toward the fundamental theorem of algebra came from calculus. The well-known method known as partial fractions for integrating a quotient of two polynomials reduces all such problems to the purely algebraic problem of factoring the denominator. It is not immediately obvious that the denominator can be factored into linear and quadratic real factors; that is the content of the fundamental theorem of algebra. Johann Bernoulli (1667-1748, the first of three mathematicians named Johann in the Bernoulli family) asserted in a paper in the *Acta eruditorum* in 1702 that such a factoring was always possible, and therefore all rational functions could be integrated. Leibniz did not agree, arguing that the polynomial $x^4 + a^2$, for example, could not be factored into quadratic factors over the reals. Here we see a great mathematician being misled by following a method. He recognized that the factorization had to be $(x^2 + a^2\sqrt{-1})(x^2 - a^2\sqrt{-1})$ and that the first factor should therefore be factored as $(x + a\sqrt{-\sqrt{-1}})(x - a\sqrt{-\sqrt{-1}})$ and the second factor as $(x + a\sqrt{\sqrt{-1}})(x - a\sqrt{\sqrt{-1}})$, but he did not realize that these factors could be combined to yield $x^4 + a^2 = (x^2 - \sqrt{2}ax + a^2)(x^2 + \sqrt{2}ax + a^2)$. It was pointed out by Niklaus Bernoulli (1687-1759, known as Niklaus I) in the *Acta eruditorum* of 1719 (three years after the death of Leibniz) that this last factorization was a consequence of the identity $x^4 + a^4 = (x^2 + a^2)^2 - 2a^2x^2$.

1.4. Euler, d'Alembert, and Lagrange. The eighteenth century saw considerable progress in the understanding of equations in general and the procedures needed to solve them. Much of this new understanding came from the two men who dominated mathematical life in that century, Euler and Lagrange.

Euler. In his 1749 paper “Recherches sur les racines imaginaires des équations” (“Investigations into the imaginary roots of equations”), devoted to equations whose degree is a power of 2 and published in the memoirs of the Berlin Academy, Euler showed that when the coefficients of a polynomial are real, its roots occur in conjugate pairs, and therefore produce irreducible real quadratic factors of the form $(x - a)^2 + b^2$. In this paper Euler argued that every polynomial of degree 2^n with real coefficients can be factored as a product of two polynomials of degree 2^{n-1} with real coefficients. In the course of the proof Euler presented the germ of an idea that was to have profound consequences. In showing that a polynomial of degree 8 could be written as a product of two polynomials of degree 4, he assumed that the coefficient of x^7 was made equal to zero by means of a linear substitution. The remaining polynomial $x^8 - ax^6 + bx^5 - cx^4 - dx^2 + ex - f$ was then to be written as a product

$$(x^4 - ux^3 + \alpha x^2 + \beta x + \gamma)(x^4 + ux^3 + \delta x^2 + \varepsilon x + \zeta).$$

Euler noted that since u was the sum of four roots of the equation, it could assume (potentially) 70 values (the number of combinations of eight things taken four at a time), and its square would satisfy an equation of degree 35.

In this paper, Euler also conjectured that the roots of an equation of degree higher than 4 cannot be constructed by applying a finite number of algebraic operations to the coefficients. This was the first explicit statement of such a conjecture.

In his 1762 paper “De resolutione aequationum cuiusque gradus” (“On the solution of equations of any degree”), published in the proceedings of the Petersburg Academy, Euler tried a different approach,³ assuming a solution of the form

$$x = w + A \sqrt[n]{v} + B \sqrt[n]{v^2} + \cdots + Q \sqrt[n]{v^{n-1}},$$

where w is a real number and v and the coefficients A, \dots, Q are to be found by a procedure resembling a Tschirnhaus transformation. This approach was useful for equations of degree 2^n , but fell short of being a general solution of the problem.

D'Alembert. Euler's contemporary and correspondent Jean le Rond d'Alembert (1717–1783) tried to prove that all polynomials could be factored into linear and quadratic factors in order to prove that all rational functions could be integrated by partial fractions. In the course of his argument he assumed that any algebraic function could be expanded in a series of fractional powers of the independent variable. While Euler was convinced by this proof, he also wrote to d'Alembert to say that this assumption would be questioned (Bottazzini, 1986, pp. 15–18).

Lagrange. In 1770 Lagrange made a survey of the methods known up to his time for solving general equations. He devoted a great deal of space to a preliminary analysis of the cubic and quartic equations. In particular, he was intrigued by the fact that the resolvent equation, which he called the *reduced* equation (*équation réduite*), for the cubic was actually an equation of degree 6 that just happened to be quadratic in the third power of the unknown. He showed that if the roots of the cubic equation $x^3 + px = q$ being solved were a , b , and c , then a root of the resolvent would be

$$y = \frac{a + \alpha b + \alpha^2 c}{3},$$

³ This approach was discovered independently by Étienne Bézout (1730–1783).

where $\alpha^3 = 1$, $\alpha \neq 1$. He argued that since the original equation was symmetric in a , b , and c , the resolvent would have to admit this y as a root, no matter how the letters a , b , and c were permuted. It therefore followed that the resolvent would in general have six different roots.

For the quartic equation with roots a , b , c , and d , he showed that the resolvent cubic equation would have a root

$$t = \frac{ab + cd}{2}.$$

Since this expression could assume only three different values when the roots were permuted—namely, half of $ab + cd$, $ac + bd$, or $ad + bc$ —it would have to satisfy an equation of degree three.

Proceeding to equations of fifth degree, Lagrange noted the only methods proposed up to that time, by Tschirnhaus and Euler-Bézout, and showed that the resolvent to be expected in all cases would be of degree 24. Pointing out that even Tschirnhaus, Euler, and Bézout themselves had not seriously attacked equations of degree five or higher, nor had anyone else tried to extend their methods, he said, "It is therefore greatly to be desired that one could estimate *a priori* the success that is to be expected in applying these methods to degrees higher than the fourth." He then set out to provide proof that, in general, one could not expect the resolvent equation to reduce to lower degree than the original equation in such cases, at least using the methods mentioned.

To prove his point, Lagrange analyzed the method of Tschirnhaus from a more general point of view. For cubic and quartic equations, in which only two coefficients needed to be eliminated (the linear and quadratic terms in the cubic, the linear and cubic terms in the quartic) the substitution $y = x^2 + ax + b$ would always work, since the elimination procedure resulted in linear and quadratic expressions in a and b in the coefficients that needed to be eliminated. Still, as Lagrange remarked, that meant two pairs of possible values (a, b) and hence really two cubic resolvents to be solved. The resolvent was therefore once again an equation of degree 6, which happened to factor into the product of two cubics. He noted what must be an ominous sign for those hoping to solve all algebraic equations by algebraic methods: The construction of the coefficients in the resolvent for an equation of degree n appeared to require solving $n - 1$ equations in $n - 1$ unknowns, of degrees $1, 2, \dots, n - 1$, so that eliminating the variable x in these equations therefore led to an expression for x that was of degree $(n - 1)!$ in y , and hence to a resolvent equation of degree $n!$ in y .

Lagrange summed up his analysis as follows:

To apply, for example, the method of Tschirnhaus to the equation of degree 5, one must solve four equations in four unknowns, the first being of degree 1, the second of degree 2, and so on. Thus the final equation resulting from the elimination of three of these unknowns will in general be of degree 24. But apart from the immense amount of labor needed to derive this equation, it is clear that after finding it, one will be hardly better off than before, unless one can reduce it to an equation of degree less than 5; and if such a reduction is possible, it can only be by dint of further labor, even more extensive than before.

The technique of counting the number of different values the root of the resolvent will have when the roots of the original equation are permuted among themselves was an important clue in solving the problem of the quintic.

1.5. Gauss and the fundamental theorem of algebra. The question of the theoretical existence of roots was settled on an intuitive level in the 1799 dissertation of Gauss. Gauss distinguished between the abstract *existence* of a root, which he proved, and an algebraic *algorithm* for finding it, the existence of which he doubted. He pointed out that attempts to prove the existence of a root and any possible algorithm for finding it must assume the possibility of extracting the n th root of a complex number. He also noted the opinion, first stated by Euler, that no algebraic algorithm existed for solving the general quintic.

The reason we say that the existence of roots was settled only on the intuitive level is that the proof of the fundamental theorem of algebra is as much topological as algebraic. The existence of real roots of an equation of odd degree with real coefficients seems obvious since a real polynomial of odd degree tends to oppositely signed infinities as the independent variable ranges from one infinity to the other. It thus follows by connectivity that it must assume a zero at some point. Gauss' proof of the existence of complex roots was similar. Much of what he was doing was new at the time, and he had to explain it in considerable detail. For that reason, he preferred to use only real-variable methods, so as not to raise any additional doubts with the use of complex numbers. In fact, he stated his purpose in that way: to prove that every equation with real coefficients has a complete factorization into linear and quadratic real polynomials.

The complex-variable background of the proof is obvious nowadays, and Gauss admitted that his lemmas were normally proved using complex numbers. The steps were as follows. First, considering the equation $z^m + Az^{m-1} + Bz^{m-2} + \cdots + Kz^2 + Lz + M = 0$, where all coefficients A, \dots, M were real numbers,⁴ taking $z = r(\cos \varphi + i \sin \varphi)$ and using the relation $z^m = r^m(\cos m\varphi + i \sin m\varphi)$, one can see that finding a root amounts to setting the real and imaginary parts equal to zero simultaneously, that is, finding r and φ such that

$$\begin{aligned} r^m \cos m\varphi + Ar^{m-1} \cos(m-1)\varphi + \cdots + Kr^2 \cos 2\varphi + Lr \cos \varphi + M &= 0, \\ r^m \sin m\varphi + Ar^{m-1} \sin(m-1)\varphi + \cdots + Kr^2 \sin 2\varphi + Lr \sin \varphi &= 0. \end{aligned}$$

What remained was to show that there actually were points where the two curves intersected. For that purpose, Gauss divided both equations by r^m and argued that for large values of r the two curves must have zeros near the zeros of $\cos m\varphi = 0$ and $\sin m\varphi = 0$. That would mean that on a sufficiently large circle, each would have $2m$ zeros, and moreover the zeros of one curve, being near the points with polar angles $(k + 1/2)\pi/m$ must separate those of the other, which are near the points with polar angles $k\pi/m$. Then, arguing that the portion of each curve inside the disk of radius r was connected, he said that it was obvious that one could not join all the pairs from one set and all the pairs from the other set using two curves that do not intersect.

Gauss was uneasy about the intuitive aspect of the proof. During his lifetime he gave several other proofs of the theorem that he regarded as more rigorous.

⁴ This restriction involves no loss of generality (see Problem 15.1).

1.6. Ruffini. As it turned out, Gauss had no need to publish his own research on the quintic equation. In the very year in which he wrote his dissertation, the first claim of a proof that it is impossible to find a formula for solving all quintic equations by algebraic operations was made by the Italian physician Paolo Ruffini (1765–1822). Ruffini's proof was based on Lagrange's count of the number of values a function can assume when its variables are permuted.⁵ The principles of such a proof were gradually coming into focus. Waring's proof that every symmetric function of the roots of a polynomial is a function of its coefficients was an important step, as was the idea of counting the number of different values a rational function of the roots can assume. To get the general proof, it was necessary to show that the root extractions performed in the course of a hypothetical solution would also be rational functions of the roots. That this is the case for quadratic and cubic equations is not difficult to see, since the quadratic formula for solving $x^2 - (r_1 + r_2)x + r_1r_2 = 0$ involves taking only one square root:

$$\sqrt{(r_1 + r_2)^2 - 4r_1r_2} = \sqrt{(r_1 - r_2)^2}.$$

Similarly, the Cardano formula for solving $y^3 + (r_1r_2 + r_2r_3 + r_3r_1)y = r_1r_2r_3$, where $r_1 + r_2 + r_3 = 0$, involves taking

$$\sqrt{\frac{(r_1r_2 + r_2r_3 + r_3r_1)^3}{27} + \frac{(r_1r_2r_3)^2}{4}} = \sqrt{\frac{-1}{108} \left((r_1 - r_2)(2r_1^2 + 5r_1r_2 + 2r_2^2) \right)},$$

followed by extraction of the cube roots of the two numbers

$$\frac{i}{3\sqrt{3}}(r_1 + \omega r_2)^2 \quad \text{and} \quad \frac{i}{3\sqrt{3}}(r_1 + \omega^2 r_2)^2,$$

where $\omega = -1/2 + i\sqrt{3}/2$ is a complex cube root of 1. These radicals are consequently rational (but not symmetric) functions of the roots.

1.7. Cauchy. Although Ruffini's proof was not generally accepted by his contemporaries, it was endorsed many years later by Augustin-Louis Cauchy (1789–1856). In 1812 Cauchy wrote a paper "Essai sur les fonctions symétriques" in which he proved the crucial fact that a function of n variables that assumes fewer values than the largest prime number less than n when the variables are permuted, actually assumes at most two values. In 1815 he published this result.

Cauchy gave credit to Lagrange, Alexandre Théophile Vandermonde (1735–1796), and Ruffini for earlier work in this area. Vandermonde, in particular, exhibited the Vandermonde determinant

$$\det \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{bmatrix} = - (x_1 - x_2)(x_1 - x_3) \cdots (x_1 - x_n)(x_2 - x_3) \cdots (x_2 - x_n) \cdots (x_{n-1} - x_n),$$

which assumes only two values, since interchanging two variables permutes the rows of the determinant and hence reverses the sign of the determinant.

⁵ An exposition of Ruffini's proof, clothed in modern terminology that Ruffini would not have recognized, can be found in the paper of Ayoub (1980).

Ruffini had shown that it was not possible to exhibit a function of five variables that could be changed into three different functions or four different functions by permuting the variables. It was this work that Cauchy proposed to generalize.

Cauchy's theorem was an elegant piece of work in the theory of finite permutation groups. To prove it, he had to invent a good deal of that theory. He pointed out that the number of permutations N equals $n!$, and that the number of those permutations that leave the function unchanged is a divisor of N , which he denoted M . In a manner now familiar, he showed that the number of different values (that is, different functions of the variables) that can be obtained by permuting the variables is $R = N/M$, and that if S is a permutation that leaves the function unchanged and T changes its value from K to K' , then ST also changes its value from K to K' . He then introduced cyclic permutations and what we now call the order of a cyclic permutation. The matrix notation now sometimes used for permutations and the notation $(\alpha\beta\gamma)$ for a permutation that maps α to β , β to γ , and γ to α , leaving all other elements fixed, was introduced in this paper.

Cauchy showed that if a permutation U is of order m , the complete set of permutations breaks up into N/m pairwise disjoint subsets (now called cosets) of m elements each. If $m > R$, which means $M > N/m$, some coset must contain two distinct elements S and T that leave the function invariant. When m is a prime p , this fact implies that some power U^s with s between 1 and $p-1$ leaves the function invariant, and since every power U^{sk} then leaves the function invariant, it follows that all powers of U leave the function invariant. If p is 2, this is not a strong statement, since $R = 1$ in that case, and all permutations whatsoever leave the function invariant. For $p > 2$, it implies that the set of permutations that leave the function invariant contains all permutations of order p .

Cauchy then showed that this set must contain all permutations of order 3, by explicitly writing any permutation of order 3 as the composition of two permutations of order p .⁶ It then followed that the permutation group can produce at most two different functions. For this case Cauchy showed that the function must be of the form $K + SV$, where K and S are symmetric and V is the Vandermonde determinant mentioned above, which switches sign when any two of its arguments are interchanged.

Besides the notation for permutations and cycles, Cauchy also invented some of the terminology of group theory, including the word *index* (*indice*) still used for the number of cosets of a subgroup of a finite group. For the number of elements M in the subgroup, he used the term *indicial* (or *indicative*) *divisor* (*diviseur indicatif*). He proposed the name *substitution* (of one permutation into another) for the composition of two permutations, and he called two permutations *equivalent* if they produce the same function, that is, they are equal modulo the subgroup of permutations that leave the function invariant. To picture cyclic permutations of finite order, he suggested arranging the distinct powers as the vertices of a regular polygon and thinking of the composition of two of them as a clockwise rotation (he said "a rotation from east to west") of the polygon. Such an arrangement suggests studying the symmetries of these polygons. However, although he frequently referred to "groups of indices" in this paper, he did not define the notion of a group in its modern sense.

⁶ The number $N = n!$ has no prime factors larger than n , so that $p \leq n$ in any case.

1.8. Abel. Cauchy's work had a profound influence on two young geniuses whose lives were destined to be very short. The first of these, the Norwegian mathematician Niels Henrik Abel (1802–1829), believed in 1821 that he had succeeded in solving the quintic equation. He sent his solution to the Danish mathematician Ferdinand Degen (1766–1825), who asked him to provide a worked-out example of a quintic equation that could be solved by Abel's method. While working through the details of an example, Abel realized his mistake. In 1824 he constructed an argument to show that such a solution was impossible and had the proof published privately. A formal version was published in the *Journal für die reine und angewandte Mathematik* in 1826. Abel was aware of Ruffini's work, and mentioned it in his argument. He attempted to fill in the gap in Ruffini's work with a proof that the intermediate radicals in any supposed solution by formula can be expressed as rational functions of the roots.

Abel's idea was that if some finite sequence of rational operations and root extractions applied to the coefficients produces a root of the equation

$$x^5 - ax^4 + bx^3 - cx^2 + dx - e = 0,$$

the final result must be expressible in the form

$$x = p + R^{\frac{1}{m}} + p_2 R^{\frac{2}{m}} + \cdots + p_{m-1} R^{\frac{m-1}{m}},$$

where p, p_2, \dots, p_{m-1} , and R are also formed by rational operations and root extractions applied to the coefficients, m is a prime number,⁷ and $R^{1/m}$ is not expressible as a rational function of the coefficients $a, b, c, d, e, p, p_2, \dots, p_{m-1}$.⁸ By straightforward reasoning on a system of linear equations for the coefficients p_j , he was able to show that R is a symmetric function of the roots, and hence that $R^{1/m}$ must assume exactly m different values as the roots are permuted. Moreover, since there are $5!$ permutations of the roots and m is a prime, it followed that $m = 2$ or $m = 5$, the case $m = 3$ having been ruled out by Cauchy. The hypothesis that $m = 5$ led to an equation in which the left-hand side assumed only five values while the right-hand side assumed 120 values as the roots were permuted. Then the hypothesis $m = 2$ led to a similar equation in which one side assumed 120 values and the other only 10. Abel concluded that the hypothesis that there exists an algorithm for solving the equation was incorrect.

The standard version of the history of mathematics credits Abel with being "the" person who proved the impossibility of solving the quintic equation. But according to Ayoub (1980, p. 274), in 1832 the Prague Scientific Society declared the proofs of Ruffini and Abel unsatisfactory and offered a prize for a correct proof. The question was investigated by William Rowan Hamilton in a report to the Royal Society in 1836 and published in the *Transactions of the Royal Irish Academy* in 1839. Hamilton's report was so heavily laden with subscripts and superscripts bearing primes that only the most dedicated reader would attempt to understand it, although Felix Klein was later (1884) to describe it as being "as lucid as it is voluminous." The proof was described by the American number theorist and historian

⁷ Extracting any root is tantamount to the sequential extraction of prime roots. Hence every root extraction in the hypothetical process of solving the equation can be assumed to be the extraction of a prime root.

⁸ Abel incorporated the apparently missing coefficient p_1 into R here, since he saw no loss of generality in doing so. A decade later, Hamilton pointed out that doing so might increase the index of the root that needed to be extracted, since p_1 might itself require the extraction of an m th root.

of mathematics Leonard Eugene Dickson as “a very complicated reconstruction of Abel’s proof.” Hamilton regarded the problem of the solvability of the quintic as still open. He wrote:

[T]he opinions of mathematicians appear to be not yet entirely agreed respecting the possibility or impossibility of expressing a root as a function of the coefficients by any finite combination of radicals and rational functions.

The verdict of history has been that Abel’s proof, suitably worded, is correct. Ruffini also had a sound method (see Ayoub, 1980), but needed to make certain subtle distinctions that were noticed only after the problem was better understood. By the end of the nineteenth century, Klein (see 1884) referred to “the proofs of *Ruffini* and *Abel*, by which it is established that a solution of the general equation of the fifth degree by extracting a finite number of roots is impossible.”

Besides his impossibility proof, Abel made positive contributions to the solution of equations. He generalized the work of Gauss on the cyclotomic (circle-splitting) equation $x^n + x^{n-1} + \cdots + x + 1 = 0$, which had led Gauss to the construction of the regular 17-sided polygon. Abel showed that if every root of an equation could be generated by applying a given rational function successively to a single (primitive) root, the equation could be solved by radicals. Any two permutations that leave this function invariant necessarily commute with each other. As a result, nowadays any group whose elements commute is called an *Abelian* group.

1.9. Galois. More light was shed on the solution of equations by the work of Abel’s contemporary Évariste Galois (1811–1832), a volatile young man who did not live to become even mature. As is well known, he died at the age of 20 in a duel fought with one of his fellow Republicans.⁹

The neatly systematized concepts of group, ring, and field that now make modern algebra the beautiful subject that it is grew out of the work of Abel and Galois, but neither of these two short-lived geniuses had a full picture of any of them. The absence of the notion of a field seems to be the most noticeable lacuna in the theorems they were proving. Where we now talk easily about *algebraic and transcendental field extensions* and regard the general equation of degree n over a field F as $x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$, where a_j is transcendental over F , Galois had to explain that the concept of a rational function was relative to what was given. For an equation with numerical coefficients, a rational function was simply a quotient of two polynomials with numerical coefficients, while if the equation had letters as coefficients, a rational function meant a quotient of two polynomials whose coefficients were rational functions of the coefficients of the equation. Even the concept of a group, which is associated with Galois, is not stated formally in any of his work. He does use the word *group* frequently in referring to a set of permutations of the roots of an equation, and he uses the properties that

⁹ The word *Republican* (*republicain*) is being used in its French sense, of course, not the American sense. It is approximately the opposite of *royaliste*. There are murky details about the duel, but it appears that the gun Galois used was not loaded, probably because he did not wish to kill a comrade-in-arms. It is also possible that the combatants had jointly decided to let fate determine the outcome and each picked up a weapon not knowing which of the two guns was loaded. The cause of the duel is also not entirely clear. The notes that Galois left behind seem to imply that he felt it necessary to warn his friends about what he considered to be the wiles of a certain young woman by whom he felt betrayed, and they felt obliged to defend her honor against his remarks.

we associate with a group: the composition of permutations. However, it is clear from his language that what makes a set of permutations a group is that *all of them have the same effect on certain rational functions of the roots*. In particular, when what we now call a group is decomposed into cosets over a subgroup, Galois refers to the cosets as groups, since any two elements of a given coset have the same effect on the rational functions. He says that a group, in this sense, may begin with any permutation at all, since there is no need to specify any natural initial order of the roots.

Besides the shortness of their lives, Abel and Galois had another thing in common: neglect of their achievements by the Paris Academy of Sciences. We shall see some details of Abel's case in Chapter 17. As for Galois, he had been expelled from the École Normale because of his Republican activities and had been in prison. He left a second paper on the subject among his effects, which was finally published in 1846.¹⁰ It had been written in January 1831, 17 months before his death, and it contained the following plaintive preface:

The attached paper is excerpted from a work that I had the honor to present to the Academy a year ago. Since this work was not understood, and doubt was cast on the propositions that it contains, I have had to settle for giving the general principles and *only one* application of my theoric in systematic order. I beg the referees at least to read these few pages with attention. [Picard, 1897, p. 33]

The language and notation used by Galois are very close to those of Lagrange. He considers an equation of degree n and claims that there exists a function (polynomial) $\varphi(a, b, c, d, \dots)$ that takes on $n!$ different values when the roots are permuted. Such a polynomial, he says, can be $\varphi(a, b, c, d, \dots) = Aa + Bb + Cc + Dd + \dots$, where A, B, C, D , and so on, are positive integers. He then fixes one root a and forms a function of two variables

$$f(V, a) = \prod (V - \varphi(a, b, c, d, \dots)),$$

(the Galois resolvent), in which the product extends over all permutations that leave a fixed. Since the function on the right is symmetric in b, c, d, \dots , all of these variables can be replaced by suitable combinations of a and the coefficients p_1, \dots, p_n (see Problem 15.4). The equation $f(\varphi(a, b, c, d, \dots), x) = 0$ then has the solution $x = a$, but has no other roots in common with the equation $p(x) = 0$. Finding the greatest common divisor of these two polynomials then makes it possible to express a as a rational function of $\varphi(a, b, c, d, \dots)$. Galois cited one of Abel's memoirs (on elliptic functions) as having stated this theorem without proof.

The main theorem of the memoir was the following: *For any equation, there is a group of permutations of the roots such that every function of the roots that is invariant under the group can be expressed rationally in terms of the coefficients of the equation, and conversely, every such function is invariant under the group.* We would nowadays say that the elements of this group generate automorphisms of the splitting field of the equation that leaves the field of coefficients invariant. As his formulation shows, Galois had only the skeleton of that result. He called the group of permutations in question the *group of the equation*. His groups are

¹⁰ Abel's great work on integrals of algebraic functions, submitted in 1827, was finally published, at the insistence of Jacobi, in 1841.

all concrete objects—permutations of the roots of equations. He developed Galois theory to the extent of analyzing what happened to the group of an equation when, in modern terms, a new element is adjoined to the base field. Galois could not be so clear. He said, “When we agree thus to regard certain quantities as known, we shall say that we are *adjoining* them to the equation being solved and that these quantities are *adjoined* to the equation.” He thought of the new element as a root of an auxiliary polynomial (the minimal polynomial of the new element, in our terms), since that is where he got the elements that he adjoined. Instead of saying that the original group might be decomposed into the cosets of the group of the new equation when all the roots of the auxiliary equation are adjoined, he said it might split into p groups, each belonging to the equation. He noted that “these groups have the remarkable property that one can pass from one to the other by operating on all the permutations with the same letter substitution.”

In a letter to a friend written the night before the duel in which he died, Galois showed that he had gone still further into this subject, making the distinction between proper and improper decompositions of the group of an equation, that is, the distinction we now make between normal and nonnormal subgroups.

Galois theory. The ideas of Galois and his predecessors were developed further by Laurent Wantzel (1814–1848) and Enrico Betti. In 1837 Wantzel used Galois’ ideas to prove that it is impossible to double the cube or trisect the angle using ruler and compass; in 1845 he proved that it is impossible to solve all equations in radicals. In 1852 Betti published a series of theorems elucidating the theory of solvability by radicals. In this way, group theory proved to be the key not only to the solvability of equations but to the full understanding of classical problems. When Ferdinand Lindemann (1852–1939) proved in 1881 that π is a transcendental number, it followed that no ruler-and-compass quadrature of the circle was possible.

The proof that the general quintic equation of degree 5 was not solvable by radicals naturally raised two questions: (1) How *can* the general quintic equation of degree 5 be solved? (2) Which particular quintic equations *can* be solved by radicals? These questions required some time to answer.

Solution of the general quintic by elliptic integrals. A partial answer to the first question came from the young mathematician Ferdinand Eisenstein (1823–1852), who showed in 1844 that the general quintic equation could be solved in terms of a function $\chi(\lambda)$ that satisfies the special quintic equation

$$(\chi(\lambda))^5 + \chi(\lambda) = \lambda,$$

This function is in a sense an analog of root extraction, since the square root function φ and the cube root function ψ satisfy the equations

$$\begin{aligned} (\varphi(\lambda))^2 &= \lambda, \\ (\psi(\lambda))^3 &= \lambda. \end{aligned}$$

Eisenstein’s solution stands somewhat apart from the main line of development, but in modern times it begins to look more reasonable. To solve all quadratic equations in a field of characteristic 2, for example, it is necessary to assume, in addition to the possibility of extracting a square root, that one has solutions to the equation $x^2 + x + 1 = 0$; these roots must be created by fiat. For a full discussion of Eisenstein’s paper, see the article of Patterson (1990).

As elliptic integrals—integrals containing the square root of a cubic or quartic polynomial—became better understood, both computational and theoretical considerations brought about a focus on transformations of one elliptic integral to another. In 1828 Jacobi studied rational changes of variable $y = U(x)/V(x)$, where U and V are polynomials of degree at most n , and found an algebraic equation that U and V must satisfy in order for this transformation to convert an elliptic integral containing one parameter (modulus) into another.

The transformation

$$\int \frac{dx}{\sqrt{(1-x^2)(1-\kappa^2 x^2)}} \mapsto \int \frac{dy}{\sqrt{(1-y^2)(1-\lambda^2 y^2)}}$$

corresponds to an equation

$$u^6 - v^6 + 5u^2v^2(u^2 - v^2) + 4uv(1 - u^4v^4) = 0,$$

where $u = \sqrt[3]{\kappa}$ and $v = \sqrt[3]{\lambda}$ (see Klein, 1884, Part II, Chapter 1, Section 3). Galois had recognized this connection and noted that the general modular equation of degree 6 could be reduced to an equation of degree 5 of which it was a resolvent. The parameter u can be expressed as a quotient of two infinite series (theta functions) in the number $q = e^{-\pi K'/K}$, where K and K' are the complete elliptic integrals of first kind with moduli κ and $\sqrt{1-\kappa^2}$. Thus, a family of equations of degree 6 containing a parameter could be solved using the elliptic modular function. It was finally Charles Hermite (1822–1901) who, in 1858, made all these facts fit together in a solution of the general quintic equation using elliptic functions.

Solution of particular quintics by radicals. The study of particular quintics that are solvable by radicals has occupied considerably more time. It is not difficult to reduce the general problem to the study of equations of the form $x^5 + px + q = 0$ via a Tschirnhaus transformation. This topic was studied by Carl Runge (1856–1927) in an 1886 paper in the *Acta mathematica*. There are only five groups of permutations of five letters that leave no letter fixed and hence could be the group of an irreducible quintic equation. They contain respectively 5, 10, 20, 60, and 120 permutations. A quintic equation having one of the first three as its group will be solvable by radicals, whereas an equation having either of the other two groups will not be. The actual construction of the solution, however, is by no means trivial. The situation is similar to that involved in the construction of regular polygons with ruler and compass. Thanks to Galois theory, we now know that it is possible for a person with sufficient patience to construct a 17-sided regular polygon—that is, partition a circle into 17 equal arcs—using ruler and compass, and Gauss actually did so.¹¹ The details of the construction, however, are quite complicated. The same theory assures us that it is similarly possible to divide the circle into 65,537 congruent arcs, a task attempted by Johann Hermes (see p. 189). In contrast, algorithms have been produced for solving quintics by radicals where it is possible to do so.¹² An early summary of results in this direction was the famous book by Felix Klein on the icosahedron (1884). An up-to-date study of the theory of solvability of equations of all degrees, with historical documentation, is the book of R. Bruce King (1996).

¹¹ Abel, using elliptic functions, partitioned the lemniscate into 17 arcs of equal length.

¹² See the paper by D. S. Dummit “Solving solvable quintics,” in *Mathematics of Computation*, 57 (1991), No. 195, 387–401.

2. Algebraic structures

The concept of a group was the first of the many abstractions that make up the world of modern algebra. We have seen how it arises through the study of the permutations of the roots of an equation. In the work of Lagrange, Ruffini, Cauchy, and Abel, only the number of different forms that a function of the roots could take was studied. Then Galois focused attention on the structure of the permutations themselves, and the result was the first abstract structure, a permutation group. Another two decades passed before the idea of a group was made abstract by Arthur Cayley (1821–1895) in 1849. Cayley defined a group as a set of symbols that could be combined in a way that was associative (he used the word) but not necessarily commutative, and such that the elements must repeat themselves if all are operated on by the same element. (From Cayley's language it is not clear whether he intended this last property as an axiom or believed that it followed from the other properties of a group.) An important example given by Cayley was a group of matrices.¹³ The complete set of axioms for an abstract group was stated by Walther von Dyck (1856–1934), a student of Felix Klein, in 1883.

2.1. Fields, rings, and algebras. The concept of a group arose in the study of the procedures used to solve equations, but that study involved other concepts that were also somewhat vague and in need of clarification. What exactly did Galois mean when he said “if we agree to regard certain objects as known” and spoke of adjoining roots to an equation? Rational *expressions* in variables representing unspecified numbers had long been part of the discourse in the solution of algebraic equations. Both Abel and Galois made frequent use of this concept. Over the course of the nineteenth century this domain of rationality evolved into what Dedekind in 1858 called a *Zahlkörper* (*number body*). Leopold Kronecker (1823–1891) preferred the term *Rationalitäts-Bereich* (*domain of rationality*). The abstract object that grew out of this concept eventually came to be known in French as a *corps*, in German as a *Körper*, and in English as a *field*.¹⁴ Dedekind considered only fields built on top of the rational numbers. Finite fields were first introduced, along with the word *field* itself, by E. H. Moore (1862–1932) in a paper published in the *Bulletin of the New York Mathematical Society* in 1893.

Other algebraic concepts arise as generalizations of number systems. In particular, the integers, the complex numbers of the form $m + ni$ (the Gaussian integers), and the integers modulo a fixed integer m led to the general concept that Hilbert, in his exhaustive 1897 report to the German Mathematical Union “Die Theorie des allgemeinen Zahlkörpers” (“The theory of the general number field”), called a *Zahlring* (*Number ring*). He gave as an equivalent term *Integritätsbereich* (*integral domain*). Both, however, were names for sets of complex numbers. Nowadays, an integral domain is defined abstractly as a commutative ring with identity in which the product of nonzero elements is nonzero; that is, there do not exist *zero divisors*. An element that is not a zero divisor is said to be *regular*. These structures were consciously abstracted and developed into the concept of an abstract *ring* in the paper “Über die Teiler der Null und die Zerlegung von Ringen” (“On zero divisors and the decomposition of rings”) by Adolf Fraenkel (1891–1965), which appeared

¹³ The word *matrix* was Cayley's invention; the word is Latin for *womb* and is used figuratively in mining to denote an ore-bearing rock. Cayley's “wombs” bore numbers rather than ore or babies.

¹⁴ A *corps*, however, is not necessarily commutative. Strictly speaking, it corresponds to what is called in English a *division ring*.

in the *Journal für die reine und angewandte Mathematik* in 1914. In his introduction Fraenkel cited the large number of particular examples as a reason for defining the abstract object. He required that the ring have at least one right identity, an element ε such that $a\varepsilon = a$ for all a , and that for at least one of the right identities every regular element should have an inverse. The English term was introduced by Eric Temple Bell (pseudonym of John Taine, 1883–1960) in a paper in the *Bulletin of the American Mathematical Society* in 1930.

Although mathematical communication was very extensive from the nineteenth century on, there was still enough difficulty due to language and transportation that British and Continental mathematicians sometimes took very different approaches to the same subject. Such was the case in algebra, where the solution of equations and abstract number theory led Continental mathematicians in one direction, at the time when British mathematicians were pursuing an abstract approach to algebra having connections with an outstanding British school of symbolic logic. For linguistic and academic reasons, the British approach also caught on in the United States, the first American foray into mathematical research. This Anglo-American algebra has been studied by Parshall (1985).

One of the first examples of this British algebra was the algebra of quaternions, invented by William Rowan Hamilton in 1843. Hamilton had been intrigued by the complex numbers since his teenage years. He questioned the meaningfulness of writing, for example $3 + \sqrt{-5}$, since this notation made it appear that two objects of different kinds—real and imaginary numbers—were being added. To rationalize this process, he took the step that seems obvious now, regarding the two numbers as ordered pairs, so that 3 is merely an abbreviation for $(3, 0)$ and $\sqrt{-5}$ an abbreviation for $(0, \sqrt{5})$, thereby algebraizing the plane. Hamilton was very much a physicist, and he saw complex multiplication, when the numbers were put in polar form $r(\cos \theta + i \sin \theta)$, as representing rotations and dilations. For him, complex addition represented all the possible translations of the plane, and complex multiplication sufficed to describe all its rotations and dilations.

Influenced by the mysticism of the poet Coleridge (1772–1834),¹⁵ whom he knew personally, he felt that great insight would be obtained if he could similarly algebraize three-dimensional space, that is, find a way to multiply triples of numbers (x, y, z) similar to the complex multiplication of pairs (x, y) . In particular, he wished to find algebraic operations corresponding to all translations and rotations of three-dimensional space. Translations were not a problem, since ordinary addition took care of them. After much reflection, during a walk in Dublin on October 10, 1843 that has become one of the most famous events in the history of mathematics, he realized that he needed a fourth quantity, since if he used one coordinate to provide a unit x , having the property that $xy = y$ and $xz = z$, the product yz would have to be expressible symmetrically in x , y , and z . When the formulas we now write as $i^2 = j^2 = k^2 = -1$, $ij = -ji = k$, $ki = -ik = j$, $jk = -kj = i$ occurred to him during this walk, he scratched them in the stone on Brougham Bridge.¹⁶ In his 1845 paper in the *Quarterly Journal* he referred to $ix + jy + kz$ as “the vector from 0 to the point x, y, z .” The word *vector* (Latin for *carrier*) occurs in this context for the first time. A quaternion thus consists of a number and a vector. Very

¹⁵ Coleridge's most famous poem, *The Rime of the Ancient Mariner*, is full of mystical uses of the numbers 3, 7, and 9.

¹⁶ In the 160 years since then, they have been effaced.

likely it was his physical intuition that led him to make this discovery. A rotation in two-dimensional space requires only one parameter for its determination: the angle of rotation. In three-dimensional space it is necessary to specify the axis of rotation by a point on the unit sphere, and then the angle of rotation, a total of three parameters. Dilations then require a fourth parameter. Thus, although quaternions were invented to describe transformations of three-dimensional space, they require four parameters to do so.

Hamilton, an excellent physicist and astronomer, worried about simply making up symbols out of his head and manipulating them. He soon found applications of them, however; and a school of his followers grew up, dedicated to spreading the lore of quaternions. By throwing away one of Hamilton's dimensions (the one that contained the unit) and using only the three symbols i , j , and k , the American mathematician Josiah Willard Gibbs (1839–1903) developed the vector calculus as we know it today, in essentially the same language that is used now. In the language of vectors quaternions can be explained easily. A quaternion is simply the formal sum of a number and a point in three-dimensional space, such as $A = a + \alpha$ or $B = b + \beta$. As Hamilton had done with complex numbers, it is possible to rationalize this seeming absurdity by regarding the number, the point, and the quaternion itself as quadruples of numbers: $a = (a, 0, 0, 0)$, $\alpha = (0, a_1, a_2, a_3)$, $A = (a, a_1, a_2, a_3)$. The familiar cross product developed by Gibbs is obtained by regarding two vectors as quaternions, multiplying them, then setting the numerical part equal to zero (projecting from four-dimensional space to three-dimensional space). Conversely, quaternion multiplication can be defined in terms of the vector (dot and cross) products: $AB = (ab - \alpha \cdot \beta) + (a\beta + b\alpha + \alpha \times \beta)$. The quaternion $\bar{A} = a - \alpha$ is the conjugate, analogous to the complex conjugate, and has the analogous property $A\bar{A} = a^2 + \alpha \cdot \alpha = a^2 + |\alpha|^2 = a^2 + a_1^2 + a_2^2 + a_3^2$. Thus $A\bar{A}$ represents the square of Euclidean distance from A to $(0, 0, 0, 0)$ and can be denoted $|A|^2$. This equation in turn shows how to divide quaternions, multiplying by the reciprocal: $1/A = (1/|A|^2)\bar{A}$. The absolute value of a quaternion has the pleasant property that $|AB| = |A||B|$.

The Harvard professor Benjamin Peirce (1808–1880) became an enthusiast of quaternions and was already lecturing on them in 1848, only a few years after their invention. Like many mathematicians before and after, he was philosophically attracted to algebra and believed it encapsulated pure thought in a way that was unique to itself. His treatise *Linear Associative Algebra* was one of the earliest treatises in this surprisingly late-arising subject.¹⁷

On the Continent algebra developed from other roots, more geometric in nature, exemplified by Grassmann's *Ausdehnungslehre*, which was described in Section 3 of Chapter 12. To some Continental mathematicians, what the British were doing did not seem sufficiently substantial. On New Year's Day 1875, Weierstrass wrote to his pupil Sof'ya Kovalevskaya that she had much more important things to learn than Hamilton's quaternions, whose algebraic foundations, he said, were of a very trivial nature. In his discussion of quaternions Klein (1926, p. 182) remarked, "It is hardly necessary to mention that the Grassmannians and the quaternionists were bitter rivals, while each of the two schools in turn split into fiercely warring subspecies." Weierstrass himself, in 1884, gave a discussion of an algebra, including

¹⁷ I say "surprisingly" because, as anyone would agree, its basic subject matter—linear equations—is much simpler than many parts of algebra that developed earlier.

the structure constants that constitute the multiplication table for the elements of a basis. Even so, the subject seems to have caught on in only a few places in Germany. At Göttingen Emmy Noether revolutionized the subject of algebras and representations of finite groups, and the concept of a Noetherian ring is now one of the basic parts of ring theory. Yet Salomon Bochner (1899–1982), who was educated in Germany and spent the first 15 years of his professional career there before coming to Princeton, recalled that the concept of an algebra was completely new to him when he first heard a young American woman lecture on it at Oxford in 1925.¹⁸

She came from Chicago and she gave at this seminar a lecture on algebras, which left all of us totally uncomprehending what it was all about. She spoke in a well-articulated, self-confident manner, but none of us had remotely heard before the terms she used, and we were lost. [Bochner, 1974, p. 832]

Bochner went on to say that a decade later he was taken aback to find a German book with the title *Algebren* (*Algebras*—it was use of the plural that Bochner found jarring). Bochner was an extremely creative and productive analyst and differential geometer. Not until the mid-1960s did he have time to ferret out Peirce's book and sit down to read it.

2.2. Abstract groups. The general theory of groups of permutations was developed in great detail in an 1869 treatise of Camille Jordan (1838–1922). This work made the importance of groups widely known. But despite Cayley's 1849 paper in group theory, the word *group* was still being used in an imprecise sense as late as 1871, in the work of two of the founders of group theory, Felix Klein and Sophus Lie (see Hawkins, 1989, p. 286). All groups were pictured concretely, as one-to-one mappings of sets. That assumption made a cancellation law $ab = ac \implies b = c$ valid automatically. For permutations of finite sets, the cancellation law implied that every element had an inverse. The corresponding inference for mappings of infinite sets is not valid, but Klein and Lie did not notice the difference at first. Lie even thought this inference could be proved. Groups as an abstract concept, characterized by the three, four, or five axioms one finds in modern textbooks, did not arrive until the twentieth century. On the abstract concept of a group Klein (1926, pp. 335–336) commented:

This abstract formulation is helpful in the construction of proofs, but not at all adapted to the discovery of new ideas and methods; on the contrary, it rather represents the culmination of an earlier development.

Klein was quick to recognize the potential of groups of transformations as a useful tool in the study of many areas of geometry and analysis. The double periodicity of elliptic functions, for example, meant that these functions were invariant

¹⁸ The woman's name was Echo Dolores Pepper, and the Mathematics Genealogy website lists her as having received the Ph.D. at the University of Chicago in 1925. Her dissertation, *Theory of Algebras over a Quasi-field* is on record, and she published at least one paper the year after receiving the degree, "Asymptotic expression for the probability of trials connected in a chain," *Annals of Mathematics*, **2** (28) (1926–27), 318–326. I have been unable to find out any more about her.

under an infinite group of translations of the plane. Klein introduced the concept of an *automorphic function*, an analytic function $f(z)$ that is invariant under a group of fractional-linear (Möbius) transformations

$$z \mapsto \frac{az + b}{cz + d}, \quad ad - bc \neq 0.$$

Both Klein and Lie made use of groups to unify many aspects of projective geometry, and Klein suggested that various kinds of geometry could be classified in relation to the groups of transformations that leave their basic objects invariant.

Lie groups. One of the most fundamental and far-reaching applications of the group concept is due to Lie, Klein's companion from 1869, when both young geometers felt like outsiders in the intensely analytic and algebraic world of Berlin mathematics. In studying surfaces in three-dimensional space, Lie and Klein naturally encountered the problem of solving the differential equations that lead to such surfaces. Lie had the idea of solving these equations using continuous transformations that leave the differential equation invariant, in analogy with what Galois had done for algebraic equations. Klein noticed an analogy between this early work of Lie and Abel's work on the solution of equations and wrote to Lie about it. Lie was very pleased at Klein's suggestion. He believed as a matter of faith in the basic validity of this analogy and developed it into the theory of Lie groups. Lie himself did not present a whole Lie group, only a portion of it near its identity element. He considered a set of one-to-one transformations indexed by n -tuples of sufficiently small real numbers in a neighborhood of $(0, 0, \dots, 0)$ in such a way that the composition of the transformations corresponded to addition of the points that indexed them. This subject was developed by Lie, Wilhelm Killing (1847–1923), Élie Cartan (1869–1951), Hermann Weyl (1885–1955), Claude Chevalley (1909–1984), Harish-Chandra (1923–1983), and others into one of the most imposing edifices of modern mathematics. A *Lie group* is a manifold in Riemann's sense that also happens to be a group, in which the group operations (multiplication and inversion) are analytic functions of the coordinates.

Lie's work is far too complicated to summarize, but we can explain his basic ideas with a simple example. The sphere S^3 in four-dimensional space can be regarded as the set of quaternions of unit norm, that is, $A = a + \alpha$ such that $|A|^2 = a^2 + |\alpha|^2 = 1$. Because $|AB| = |A||B|$, this set is closed under quaternion multiplication and inverses. But this sphere is also a three-dimensional manifold and can be parameterized by, say, the stereographic projection from $(-1, 0, 0, 0)$ through the equatorial hyperplane consisting of points $(0, x, y, z)$. This projection maps $(0, x, y, z)$ to

$$\left(\frac{1 - x^2 - y^2 - z^2}{1 + x^2 + y^2 + z^2}, \frac{2x}{1 + x^2 + y^2 + z^2}, \frac{2y}{1 + x^2 + y^2 + z^2}, \frac{2z}{1 + x^2 + y^2 + z^2} \right).$$

This parameterization covers the entire group except for the point $(-1, 0, 0, 0)$. To parameterize a portion of the sphere containing this point requires a second parametrization, which can be projection from the opposite pole $(1, 0, 0, 0)$. When the points in the group with coordinates (u, v, w) and (x, y, z) are multiplied, the result is the point whose first coordinate is

$$\frac{-(u^2x + (-1 + v^2 + w^2)x + 2wy - 2vz + u(-1 + x^2 + y^2 + z^2))}{1 - 2(ux + vy + wz) + (u^2 + v^2 + w^2)(x^2 + y^2 + z^2)}.$$

Since we have no need to do any computations, we omit the other two coordinates. The point to be noticed is that this function is *differentiable*, so that the group operation, when interpreted in terms of the parameters, is a differentiable operation.

To study a Lie group, one passes to the tangent spaces it has as a manifold: in particular, the tangent space at the group identity. This tangent space is determined by the directions in the parameter space around the point corresponding to the identity. Each direction gives a directional derivative that operates on differentiable functions. For that reason, the tangent space is defined to be the set of differential operators of the form $X = \sum a_j(x) \frac{\partial}{\partial x_j}$. The composition of two such operators involves second derivatives, so that XY is not in general an element of the tangent space. However, the second partial derivatives cancel in the expression $XY - YX$ (the Lie bracket). This multiplication operation makes the tangent space into a *Lie algebra*. This algebra, being a finite-dimensional vector space, is determined as an algebra once the multiplication table for the elements of a basis is given. To take the simplest nontrivial example, the Lie group of rotations of three-dimensional space (represented as 3×3 rotation matrices) has the vector algebra developed by Gibbs (with the cross product as multiplication) as its Lie algebra.

Elements of the group can be generated from the Lie algebra by applying a mapping called the exponential mapping from the Lie algebra into the Lie group. Finding this mapping amounts to solving a differential equation. The resulting combination of algebra, geometry, and analysis is both profound and beautiful.

It would have been pleasant for mathematics in general and for Lie in particular if this beauty and profundity had been recognized immediately. Unfortunately, Lie's work was not well understood at first. In January 1884 he wrote to his friend, the German mathematician Adolf Mayer (1839-1908):

I am so certain, so absolutely certain, that these theories will be acknowledged as fundamental one day in the future. If I wish to procure such an opinion *soon*, it is... because I could produce ten times more. [Engel, 1899, quoted by Parshall, 1985, p. 265]

Lie's vision was soon vindicated. By the end of his life, the potential of the theory was being recognized, and its development has never slowed in the century that has elapsed since that time.

Group representations. The road to abstraction is a two-way street. Once an axiomatic characterization of an object is stated, a classification program starts automatically, aimed at answering two important questions: First, how abstract is the abstract object, really? Second, how many abstractly different objects fit the axioms?

The first question leads to the search for concrete representations of abstract groups given only by a multiplication table. We know, for example, that every group G can be thought of as a group of one-to-one mappings by associating with each $a \in G$ the mapping $L_a : G \rightarrow G$ given by $z \mapsto az$. This fact was noted by Cayley when he introduced the abstract concept. It is also easy to show that any finite group can be represented in a trivial way as a group of invertible matrices whose entries are all zeros and ones. First regard the group as a group of permutations of a set of k objects. Then make the k objects into the basis of a vector space and associate with each permutation the matrix of the linear transformation it

defines. Essentially, this representation was introduced by the American logician-philosopher Charles Sanders Peirce in 1879.

An early prefiguration of an important concept in the representation of groups occurred in an 1837 paper of Dirichlet proving that an arithmetic progression whose first term and difference are relatively prime contains infinitely many primes. As discussed in Section 1 of Chapter 8, that paper contains the *Dirichlet character* $\chi(n)$, defined as $(-1)^{(n-1)/2}$ if n is odd and 0 if n is even. This character has the property that $\chi(mn) = \chi(m)\chi(n)$. The definition of a character as a homomorphism into the multiplicative group of nonzero complex numbers was given by Dedekind in an 1879 supplement to Dirichlet's lectures. About the same time, Sylvester was showing how matrices could be used to represent the quaternions.¹⁹

The theory of representations of finite groups was developed by the German mathematician Ferdinand Georg Frobenius (1849–1917), responding to a question posed by Dedekind. The original question was simply to factor the determinant of a certain matrix associated with a finite group. In trying to solve this problem, Frobenius introduced the idea of a representation and a character of a representation. Although the subject is too technical for details to be given here, the characters, being computable, reveal certain facts about the structure of a finite group, and in some cases determine it completely.

This theory was extended to Lie groups in 1927 by Hermann Weyl and his student F. Peter. In that context, it turned out, representation theory subsumes and unifies the theories of Fourier series and Fourier integrals, both of which are ways of analyzing functions defined on a Lie group (the circle or line) by transferring them to functions defined on a separate (dual) group. The subject is now called *abstract harmonic analysis*.

Finite groups. The subject of finite groups grew up in connection with the solution of equations, as we have already seen. For that purpose, one of the most important questions was to decide which groups corresponded to equations that are solvable. Such a group G has a chain of normal subgroups $G \supset G_1 \supset G_2 \supset \cdots \supset \{1\}$ in which each factor group G_i/G_{i+1} is commutative. Because of the connection with equations, a group having such a chain of subgroups is said to be *solvable*. A solvable group can be built up from the simplest type of group, the group of integers modulo a prime, and so its structure may be regarded as known. It would be desirable to have a classification that can be used to break down any finite group into its simplest elements in a similar way. The general problem is so difficult that it is nowhere near solution. However, a significant piece of the program has been achieved: the classification of simple groups. A simple group is one whose only normal subgroups are itself and $\{1\}$.

The project of classifying these groups was referred to by one of its leaders, Daniel Gorenstein (1923–1992), as the Thirty Years' War, since a strategy for the classification was suggested by Richard Brauer (1901–1977) at the International Congress of Mathematicians in 1954 and the classification was completed in the

¹⁹ Cayley, Peirce, and Sylvester were well acquainted personally and professionally with one another. Cayley and Peirce were at Johns Hopkins University during part of the time that Sylvester was chair of mathematics there. They formed the strong core of the Anglo-American school of abstract algebra described by Parshall (1985). At the heart of this abstraction, at least in the case of Peirce, was a philosophical program of creating a universal symbolic algebra that could be applied in any situation. Such a program required that the symbols be mere symbols until applied to some specific situation.

1980s with the discovery of the last “sporadic” group. The project consists of so many complicated parts that the process of streamlining and clarifying it, with all the new projects that process will no doubt spawn, is likely to continue for many decades to come.

An important part of the project was the 1963 proof by Walter Feit and John Thompson (b. 1932) that all finite simple groups have an even number of elements. The proof was 250 pages long and occupied an entire issue of the *Pacific Journal of Mathematics*. As it turned out, this project was destined to generate large numbers of long papers. In fact, the mathematician who contributed the last step in the classification later wrote, “At least 3,000 pages of mathematically dense preprints appeared in the years 1976–1980 and simply overwhelmed the digestive system of the group theory community” (Solomon, 1995, p. 236). The outcome of the project is intensely satisfying from an aesthetic point of view. It turns out that there are three (or four) infinite classes of finite simple groups: (1) groups of prime order; (2) the group of even permutations on n letters ($n \geq 5$);²⁰ (3) certain finite linear groups, a class that can be subdivided into classical matrix groups and twisted groups of Lie type, whose exact definition is not important for present purposes. Outside those classes are the “sporadic” groups.

If this classification seems to resemble the old classification of constructions as planar, solid, and curvilinear, in which the final class is merely a catchall term for anything that doesn’t fit into the other classes, that impression is misleading. The class of sporadic groups turns out to contain precisely 26 groups, whose properties have been tabulated. The smallest of them is M_{11} with 7920 elements, one of five sporadic simple groups discovered by Émil Mathieu (1835–1890). The largest, officially denoted F_1 , is informally known as the Monster, since it consists of

$$2^{46} \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^2 \cdot 13^3 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71$$

elements. It was constructed in 1980.

2.3. Number systems. Rings and fields can be regarded as generalized number systems, since they admit addition, subtraction, and multiplication, and sometimes division as well. As noted above, such general systems have been used since Gauss began the study of arithmetic modulo an integer and proved that the factorization of a Gaussian integer $m + ni$ into irreducible Gaussian integers is unique up to a power of i . Gauss was particularly interested in these numbers, since when they are introduced, 2 is not a prime number ($2 = (1 + i)(1 - i)$), nor is any number of the form $4n + 1$ prime. For example, $5 = (2 + i)(2 - i)$. The number 3, however, remains prime. If we pass to numbers of the form $m + n\sqrt{-2}$, factorization is still unique, but this uniqueness is lost for numbers of the form $m + n\sqrt{-3}$, since $4 = 2 \cdot 2 = (1 + \sqrt{-3})(1 - \sqrt{-3})$. The Gaussian integers were the first of an increasingly abstract class of structures on which multiplication is defined and obeys a cancellation law (that is $ac = bc$ and $c \neq 0$ implies that $a = b$), but division is not necessarily always possible. If the factorization of each element is essentially unique, such a structure is called a *Gaussian domain*.

²⁰ The fact that this group is simple and noncommutative implies that the symmetric group consisting of all permutations on n letters cannot be solvable for $n \geq 5$, and hence that the general equation of degree n is not solvable by radicals.

Ideal complex numbers. Numbers of the form $m + n\omega$, where $\omega = -1/2 + \sqrt{-3}/2$, a primitive cube root of unity, satisfies the equation $\omega^2 = -1 - \omega$, have properties similar to the Gaussian integers. For this system the number 3 is not a prime, since $3 = (2 + \omega)(1 - \omega)$. Numbers of the form $m + n\omega$ do have a unique factorization into primes, but Ernst Eduard Kummer (1810–1893) discovered in 1844 that “complex integers” of the form $m + n_1\alpha + n_2\alpha^2 + \cdots + n_{\lambda-2}\alpha^{\lambda-2}$, where $1 + \alpha + \cdots + \alpha^{\lambda-1} = 0$ and λ is prime, do not necessarily have the property of unique factorization.²¹ This fact seems to have a connection with Fermat’s last theorem, which can be stated by saying that the number

$$x^\lambda + y^\lambda = (x + y)(x + \alpha y)(x + \alpha^2 y) \cdots (x + \alpha^{\lambda-1} y)$$

is never equal to z^λ for any nonzero integers x, y, z . For that reason, Klein (1926, p. 321) asserted that it was precisely in this context that Kummer made the discovery. But Edwards (1977, pp. 79–81) argues convincingly that the discovery was connected with the search for higher reciprocity laws, analogs of the quadratic reciprocity discussed in Section 1 of Chapter 8.

However that may be, what Kummer made of the discovery is quite interesting. He introduced “ideal” numbers that would divide some of the otherwise irreducible numbers, just as the imaginary number $2 + i$ divides 5, and (he said) just as one introduces ideal chords to be held in common by two circles that do not intersect in the ordinary sense. By his definition, a prime number p that equals 1 modulo λ has $\lambda - 1$ factors.²² These factors may be actual complex integers of the form stated. For example, when $\lambda = 3$, $p = 13$, we have $13 = (4 + \omega)(3 - \omega)$. If they were not actual complex numbers, he assigned an ideal factor of p to correspond to each root ξ of the congruence $\xi^\lambda \equiv 1 \pmod{p}$, thereby obtaining $\lambda - 1$ nonunit factors. His rationale was that if $f(\alpha)$ divided p in the ordinary sense, then $f(\xi)$ would be divisible by p in the ordinary sense. More generally, a complex integer $\Phi(\alpha)$ was to have the ideal factor corresponding to ξ if $\Phi(\xi) \equiv 0 \pmod{p}$. Then if $\Phi(\alpha)$ was divisible by p , it would be divisible by all of the factors of p , whether actual complex numbers or ideal complex numbers. When these ideal complex numbers were introduced, unique factorization was restored.

Questions and problems

15.1. Prove that if every polynomial with real coefficients has a zero in the complex numbers, then the same is true of every polynomial with complex coefficients. To get started, let $p(z) = z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n$ be a polynomial with complex coefficients a_1, \dots, a_n . Consider the polynomial $q(z)$ of degree $2n$ given by $q(z) = p(z)p(\bar{z})$, where the overline indicates complex conjugation. This polynomial has real coefficients, and so by hypothesis has a complex zero z_0 .

15.2. Formulate Cauchy’s 1812 result as the following theorem and prove it: *Let p be a prime number, $3 \leq p \leq n$. If a subgroup of the symmetric group on n letters contains all permutations of order p , it is either the entire symmetric group or the alternating group.*

²¹ The first prime for which unique factorization fails is $p = 23$, just in case the reader was hoping to see an example.

²² This relation between the primes p and λ speaks in favor of Edwards’ argument that Kummer’s goal had been a higher reciprocity law.

15.3. Cauchy's theorem that every cycle of order 3 can be written as the composition of two cycles of order m if $m > 3$ looks as if it ought to apply to cycles of order 2 also. What goes wrong when you try to prove this "theorem"?

15.4. Let $S_j(a, b, c, d)$ be the j th elementary symmetric polynomial, that is, the sum of all products of j distinct factors chosen from $\{a, b, c, d\}$. Prove that $S_j(a, b, c, d) = S_j(b, c, d) + aS_{j-1}(b, c, d)$. Derive as a corollary that given a polynomial equation $x^4 - S_1(a, b, c, d)x^3 + S_2(a, b, c, d)x^2 + S_3(a, b, c, d)x + S_4(a, b, c, d) = 0 = x^4 - p_1x^3 + p_2x^2 - p_3x + p_4$ having a, b, c, d as roots, each elementary symmetric function in b, c, d can be expressed in terms of a and the coefficients p_j : $S_1(b, c, d) = p_1 - a$, $S_2(b, c, d) = p_2 - aS_1(b, c, d) = p_2 - ap_1 + a^2$, $S_3(b, c, d) = p_3 - ap_2 + a^2p_1 - a^3$.

15.5. Prove that if z is a prime in the ring obtained by adjoining the p th roots of unity to the integers (where p is a prime), the equation

$$z^p = x^p + y^p$$

can hold only if $x = 0$ or $y = 0$.

15.6. Consider the complex numbers of the form $z = m + n\omega$, where $\omega = -1/2 + \sqrt{-3}/2$ is a cube root of unity. Show that $N(z) = m^2 - mn + n^2$ has the property $N(zw) = N(z)N(w)$ and that $N(z + w) \leq 2(N(z) + N(w))$. Then show that a Euclidean algorithm exists for such complex numbers: Given z and $w \neq 0$, there exist q and r such that $z = qw + r$ where $N(r) < N(w)$. Thus, a Euclidean algorithm exists for these numbers, and so they must exhibit unique factorization. [Hint: $N(z) = |z|^2$. Show that for every complex number u there exists a number q of this form such that $|q - u| < 1$. Apply this fact with $u = z/w$ and define r to be $z - qw$.]

15.7. Show that in quaternions the equation $X^2 + r^2 = 0$, where r is a positive real number (scalar), is satisfied precisely by the quaternions $X = x + \xi$ such that $x = 0$, $|\xi| = r$, that is, by all the points on the sphere of radius r . In other words, in quaternions the square roots of negative numbers are simply the nonzero vectors in three-dimensional space. Thus, even though quaternions act "almost" like the complex numbers, the absence of a commutative law makes a great difference when polynomial algebra is considered. A linear equation can have only one solution, but a quadratic equation can have an uncountable infinity of solutions.

Part 6

Analysis

The great watershed in the history of mathematics is the invention of the calculus. It synthesized nearly all the algebra and geometry that had come before and generated problems that led to much of the mathematics studied today. Although calculus is an amalgam of algebra and geometry, it soon developed results that were indispensable in other areas of mathematics. Even theories whose origins seem to be independent of all forms of geometry—combinatorics, for example—turn out to involve concepts such as generating functions, for which the calculus is essential.

Elements of the calculus had existed from the earliest times in the form of infinitesimal methods in geometry, and such techniques were refined in the early seventeenth century. Although strict boundaries in history tend to be artificial constructions, there is such a boundary between analytic geometry and calculus. That boundary is the introduction of infinitesimal methods. As approximate, easy-to-remember formulas, we can write: algebra + geometry = analytic geometry, and analytic geometry + infinitesimals = calculus.

The introduction of infinitesimals into a geometry that had only recently struggled back to the level of rigor achieved by Archimedes raised alarms in certain quarters, but the new methods led to spectacular advances in theoretical physics and geometry that have continued to the present time. Like the Pythagoreans, modern mathematicians faced the twin challenges of extending the range of applicability of their mathematics while making it more rigorous. The responses to these challenges led to the modern subject of analysis. Starting from a base of real numbers, represented as ratios of line segments and written in the symbolic language of modern algebra, mathematicians extended their formulas to complex numbers, opening up a host of new applications and creating the beautiful subject of analytic function theory (complex analysis). At the same time, they were examining the hidden assumptions in their methods and making their limiting processes more rigorous by introducing the appropriate definitions of integrals, derivatives, and series, leading to the subject of functions of a real variable (real analysis).

Part 6 consists of two chapters. The creation of calculus and its immediate outgrowths, differential equations and the calculus of variations, is described in Chapter 16, while the further development of these subjects into modern analysis is the theme of Chapter 17.

CHAPTER 16

The Calculus

The infinite occurs in three forms in calculus: the derivative, the integral, and the power series. Integration, in the form of finding areas and volumes, was developed as a particular theory before the other two subjects came into general use. As we have seen, infinitesimal methods were used in geometry by the Chinese and Japanese, and the latter also used infinite series to solve geometric problems. In India also, mathematicians used infinite series to solve geometric problems via trigonometry. According to Rajagopal (1993), the mathematician Nilakanta, who lived in South India and whose dates are given as 1444–1543, gave a general proof of the formula for the sum of a geometric series. The most advanced of these results is attributed to Madhava (1340–1425), but is definitively stated in the work of Jyeshthadeva (1530–ca. 1608):

The product of the given Sine and the radius divided by the Cosine is the first result. From the first,...etc., results obtain...a sequence of results by taking repeatedly the square of the Sine as the multiplier and the square of the Cosine as the divisor. Divide ... in order by the odd numbers one, three, etc... From the sum of the odd terms, subtract the sum of the even terms. [The result] becomes the arc. [Rajagopal, 1993, p. 98]

These instructions give in words an algorithm that we would write as the following formula, remembering that the Sine and Cosine used in earlier times correspond to our $r \sin \theta$ and $r \cos \theta$, where r is the radius of the circle:

$$r\theta = \frac{r^2 \sin \theta}{r \cos \theta} - \frac{r^4 \sin^3 \theta}{3r^3 \cos^3 \theta} + \frac{r^6 \sin^5 \theta}{5r^5 \cos^5 \theta} - \cdots$$

The bulk of calculus was developed in Europe during the seventeenth century, and it is on that development that the rest of this chapter is focused.

Since analytic geometry was discussed in Section 1 of Chapter 12, we take up the story at the point where infinitesimal methods begin to be used in finding tangents and areas. The crucial step is the realization of the mutually inverse nature of these two processes and their consolidation as a set of algebraic and limit operations that can be applied to any function. At the center of the entire process lies the very concept of a function, which was a seventeenth-century innovation.

1. Prelude to the calculus

In his comprehensive history of the calculus (1949), Boyer described “a century of anticipation” during which the application of algebra to geometric problems began to incorporate some of the less systematic parts of ancient geometry, especially the infinitesimal ideas contained in what was called the method of indivisibles. Let us

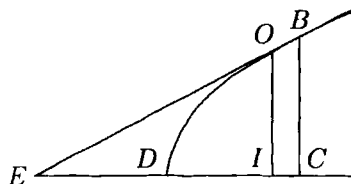


FIGURE 1. Fermat's method of finding the subtangent.

take up the story of calculus at the point where algebra enters the picture, beginning with some elementary problems of finding tangents and areas.

1.1. Tangent and maximum problems. The main problem in finding a tangent to a curve at a given point is to find some second condition, in addition to passing through the point, that this line must satisfy so as to determine it uniquely. It suffices to know either a second point that it must pass through or the angle that it must make with a given line. Fermat had attacked the problem of finding maxima and minima of variables even before the publication of Descartes' *Géométrie*. As his works were not published during his lifetime but only circulated among those who were in a rather select group of correspondents, his work in this area was not recognized for some time. His method is very close to what is still taught in calculus books. The difference is that whereas we now use the derivative to find the slope of the tangent line, that is, the tangent of the angle it makes with a reference axis, Fermat looked for the point where the tangent intercepted that axis. If the two lines did not intersect, obviously the tangent was easily determined as the unique parallel through the given point to the given axis. In all other cases Fermat needed to determine the length of the projection of the tangent on the axis from the point of intersection to the point below the point of tangency, a length known as the *subtangent*. In a letter sent to Mersenne and forwarded to Descartes in 1638 Fermat explained his method of finding the subtangent.

In Fig. 1 the curve DB is a parabola with axis CE , and the tangent at B meets the axis at E . Since the parabola is convex, a point O between B and E on the tangent lies outside the parabola. That location provided Fermat with two inequalities, one of which was $\overline{CD} : \overline{DI} > \overline{BC}^2 : \overline{OI}^2$. (Equality would hold here if \overline{OI} were replaced by the portion of it cut off by the parabola.) Since $\overline{BC} : \overline{OI} = \overline{CE} : \overline{EI}$, it follows that $\overline{CD} : \overline{DI} > \overline{CE}^2 : \overline{EI}^2$. Then abbreviating by setting $\overline{CD} = g$, $\overline{CE} = x$, and $\overline{CI} = y$, we have $g : g - y > x^2 : x^2 + y^2 - 2xy$, and cross-multiplying,

$$gx^2 + gy^2 - 2gxy > gx^2 - x^2y.$$

Canceling the term gx^2 and dividing by y , we obtain $gy - 2gx > -x^2$. Since this inequality must hold for all y (no matter how small), it follows that $x^2 \geq 2gx$, that is, $x \geq 2g$ if $x > 0$. Choosing a point O beyond B on the tangent and reasoning in the same way would give $x \leq 2g$, so that $x = 2g$. Since x was the quantity to be determined, the problem is solved.

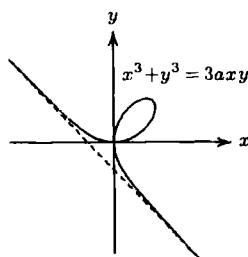


FIGURE 2. The folium of Descartes. Descartes and Fermat considered only the loop in this curve.

In this paper Fermat asserted, “And this method never fails. . . .” This assertion provoked an objection from Descartes,¹ who challenged Fermat with the curve of Fig. 2, now known as the folium of Descartes, having equation $x^3 + y^3 = 3axy$.

Descartes did not regard curves such as the spiral and the quadratrix as admissible in argument, since they are generated by two motions whose relationship to each other cannot be determined exactly. A few such curves, however, were to prove a very fruitful source of new constructions and applications. One of them, which had first been noticed in the early sixteenth century by an obscure mathematician named Charles Bouvelles (ca. 1470–ca. 1553), is the cycloid, the curve generated by a point on a circle (called the generating circle) that rolls without slipping along a straight line. We have already mentioned this curve, assuming that the reader will have heard of it, in Section 3 of Chapter 12. It is easily pictured by imagining a painted spot on the rim of a wheel as the wheel rolls along the ground. Since the linear velocity of the rim relative to its center is exactly equal to the linear velocity of the center, it follows that the point is at any instant moving along the bisector of the angle formed by a horizontal line and the tangent to the generating circle. In this way, given the generating circle, it is an easy matter to construct the tangent to the cycloid. This result was obtained independently around 1638 by Descartes, Fermat, and Gilles Personne de Roberval (1602–1675), and slightly later by Evangelista Torricelli (1608–1647), a pupil of Galileo Galilei (1564–1642).

1.2. Lengths, areas, and volumes. Seventeenth-century mathematicians had inherited two conceptually different ways of applying infinitesimal ideas to find areas and volumes. One was to regard an area as a “sum of lines.” The other was to approximate the area by a sum of regular figures and try to show that the approximation got better as the individual regular figures got smaller. The rigorous version of the latter argument, the method of exhaustion, was tedious and of limited application.

Cavalieri’s principle. In the “sum of lines” approach, a figure whose area or volume was required was sliced into parallel sections, and these sections were shown to be equal or proportional to corresponding sections of a second figure whose area or

¹ There was little love lost between Descartes and Fermat, since Fermat had dismissed Descartes’ derivation of the law of refraction. (Descartes assumed that light traveled faster in denser media; Fermat assumed that it traveled slower. Yet they both arrived at the same law! For details, see Indorato and Nastasi, 1989.) Descartes longed for revenge, and even though he eventually ended the controversy over Fermat’s methods with the equivalent of, “You should have said so in the first place, and we would never have argued. . . ,” he continued to attack Fermat’s construction of the tangent to a cycloid.

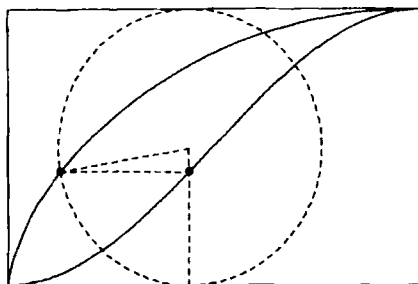


FIGURE 3. Roberval's quadrature of the cycloid.

volume was known. The first figure was then asserted to be equal or proportional to the second. The principle was stated in 1635 by Bonaventura Cavalieri (1598–1647), a Jesuit priest and a student of Galileo. At the time it was customary for professors to prove their worthiness for a chair of mathematics by a learned dissertation. Cavalieri proved certain figures equal by pairing off congruent sections of them, in a manner similar to that of Archimedes' *Method* and the method by which Zu Chongzhi and Zu Geng found the volume of a sphere. This method implied that figures in a plane lying between two parallel lines and such that all sections parallel to those lines have the same length must have equal area. This principle is now called *Cavalieri's principle*. The idea of regarding a two-dimensional figure as a sum of lines or a three-dimensional figure as a sum of plane figures was extended by Cavalieri to consideration of the squares on the lines in a plane figure, then to the cubes on the lines in a figure, and so on.

The cycloid. Cavalieri's principle was soon applied to find the area of the cycloid. Roberval, who found the tangent to the cycloid, also found the area beneath it by a clever use of the method of indivisibles. He considered along with half an arch of the cycloid itself a curve he called the *companion* to the cycloid. This companion curve is generated by a point that is always directly below or above the center of the generating circle as it rolls along and at the same height as the point on the rim that is generating the cycloid. As the circle makes half a revolution (see Fig. 3), the cycloid and its companion first diverge from the ground level, then meet again at the top. Symmetry considerations show that the area under the companion curve is exactly one-half of the rectangle whose vertical sides are the initial and final positions of the diameter of the generating circle through the point generating the cycloid. But by definition of the two curves their generating points are always at the same height, and the horizontal distance between them at any instant is half of the corresponding horizontal section of the generating circle. Hence by Cavalieri's principle the area between the two curves is exactly half the area of the circle.

Rectangular approximations and the method of exhaustion. Besides the method of indivisibles (Cavalieri's principle), mathematicians of the time also applied the method of polygonal approximation to find areas. In 1640 Fermat wrote a paper on quadratures in which he found the areas under certain figures by a method that he saw could easily be generalized. He considered a "general hyperbola," as in Fig. 4, a curve referred to asymptotes AR and AC and defined by the property

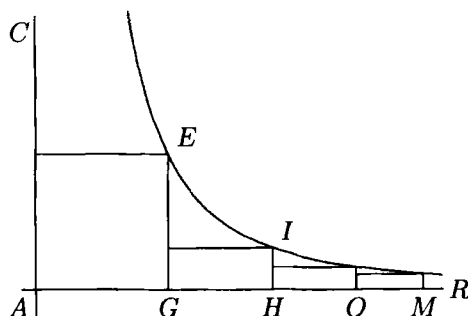


FIGURE 4. Fermat's quadrature of a generalized hyperbola.

that the ratio $AH^m : AG^m = EG^n : HI^n$ is the same for any two points E and I on the curve; we would describe this property by saying that $x^m y^n = \text{const.}$

Powers of sines. Cavalieri found the "sums of the powers of the lines" inside a triangle. In 1659 Pascal did the same for the "sums of the powers of the lines inside a quadrant of a circle." Now a line inside a quadrant of a circle is what up to now has been called a sine. Thus, Pascal found the sum of the powers of the sines of a quadrant of a circle. In modern terms, where Cavalieri found $\int_0^a x^n dx = a^{n+1}/(n+1)$, Pascal found $\int_\alpha^\beta (R \sin \varphi) R d\varphi = R(R \cos \alpha - R \cos \beta)$.

1.3. The relation between tangents and areas. The first statement of a relation between tangents and areas appears in 1670 in a book entitled *Lectiones geometricae* by Isaac Barrow (1630–1677), a professor of mathematics at Cambridge and later chaplain to Charles II. Barrow gave the credit for this theorem to "that most learned man, Gregory of Aberdeen" (James Gregory, 1638–1675). Barrow states several theorems resembling the fundamental theorem of calculus. The first theorem (Section 11 of Lecture 10) is the easiest to understand. Given a curve referred to an axis, Barrow constructs a second curve such that the ordinate at each point is proportional to the area under the original curve up to that point. We would express this relation as $F(x) = (1/R) \int_a^x f(t) dt$, where $y = f(x)$ is the first curve, $y = F(x)$ is the second, and $1/R$ is the constant of proportionality. If the point $T = (t, 0)$ is chosen on the axis so that $(x - t) \cdot f(x) = RF(x)$, then, said Barrow, T is the foot of the subtangent to the curve $y = F(x)$; that is, $x - t$ is the length of the subtangent. In modern language the length of the subtangent to the curve $y = F(x)$ is $|F(x)/F'(x)|$. This expression would replace $(x - t)$ in the equation given by Barrow. If both $F(x)$ and $F'(x)$ are positive, this relation really does say that $f(x) = RF'(x) = (d/dx) \int_a^x f(t) dt$.

Later, in Section 19 of Lecture 11, Barrow shows the other version of the fundamental theorem, that is, that if a curve is chosen so that the ratio of its ordinate to its subtangent (this ratio is precisely what we now call the derivative) is proportional to the ordinate of a second curve, the area under the second curve is proportional to the ordinate of the first.

1.4. Infinite series and products. The methods of integration requiring the summing of infinitesimal rectangles or all the lines inside a plane figure led naturally to the consideration of infinite series. Several special series were known by the mid-seventeenth century. For example, the Scottish mathematician James Gregory

published a work on geometry in 1668 in which he stated the equivalent of the formula given earlier (unknown to Gregory, of course) by Jyeshtadeva:

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots$$

Similarly, infinite product expansions were known by this time for the number π . One, due to Wallis, is

$$\frac{2}{\pi} = \frac{1 \cdot 3 \cdot 3 \cdot 5 \cdot 5 \cdot 7 \cdots}{2 \cdot 2 \cdot 4 \cdot 4 \cdot 6 \cdot 6 \cdots}$$

The binomial series. It was the binomial series that really established the use of infinite series in analysis. The expansion of a power of a binomial leads to finite series when the exponent is a nonnegative integer and to an infinite series otherwise. This series, which we now write in the form

$$(1+x)^r = 1 + \sum_{k=1}^{\infty} \frac{r(r-1) \cdots (r-k+1)}{1 \cdots k} x^k,$$

was discovered by Isaac Newton (1642–1727) around 1665, although, of course, he expressed it in a different language, as a recursive procedure for finding the terms. In a 1676 letter to Henry Oldenburg (1615–1677), the Secretary of the Royal Society, Newton wrote this expansion as

$$\overline{P+PQ} \Big| \frac{m}{n} = P \Big| \frac{m}{n} + \frac{m}{n} A Q + \frac{m-n}{2n} B Q + \frac{m-2n}{3n} C Q + \frac{m-3n}{4n} D Q + \text{etc.}$$

“where $P+PQ$ stands for a quantity whose root or power or whose root of a power is to be found, P being the first term of that quantity, Q being the remaining terms divided by the first term and m/n the numerical index of the powers of $P+PQ \dots$ A stands for the first term $P \Big| \frac{m}{n}$, B for the second term $\frac{m}{n} A Q$, and so on. . .”

Newton’s explanation of the meaning of the terms A, B, C, \dots , means that the k th term is obtained from its predecessor via multiplication by $\{[(m/n) - k]/(k+1)\}Q$. He said that m/n could be any fraction, positive or negative.

2. Newton and Leibniz

The results we have just examined show that parts of the calculus were recognized by the mid-seventeenth century, like the pieces of a jigsaw puzzle lying loose on a table. What was needed was someone to see the pattern and fit all the pieces together. The unifying principle was the concept of a derivative, and that concept came to Newton and Leibniz independently and in slightly differing forms.

2.1. Isaac Newton. Isaac Newton discovered the binomial theorem, the general use of infinite series, and what he called the *method of fluxions* during the mid-1660s. His early notes on the subject were not published until after his death, but a revised version of the method was expounded in his *Principia*.

Newton's first version of the calculus. Newton first developed the calculus in what we would call parametric form. Time was the universal independent variable, and the relative rates of change of other variables were computed as the ratios of their absolute rates of change with respect to time. Newton thought of variables as moving quantities and focused attention on their velocities. He used the letter o to represent a small time interval and p for the velocity of the variable x , so that the change in x over the time interval o was op . Similarly, using q for the velocity of y , if y and x are related by $y^n = x^m$, then $(y + oq)^n = (x + op)^m$. Both sides can be expanded by the binomial theorem. Then if the equal terms y^n and x^m are subtracted, all the remaining terms are divisible by o . When o is divided out, one side is $nqy^{n-1} + oA$ and the other is $mpx^{m-1} + oB$. Ignoring the terms containing o , since o is small, one finds that the *relative* rate of change of the two variables, q/p is given by $q/p = (mx^{m-1})/(ny^{n-1})$; and since $y = x^{m/n}$, it follows that $q/p = (m/n)x^{(m/n)-1}$. Here at last was the concept of a derivative, expressed as a relative rate of change.

Newton recognized that reversing the process of finding the relative rate of change provides a solution of the area problem. He was able to find the area under the curve $y = ax^{m/n}$ by working backward.

Fluxions and fluents. Newton's "second draft" of the calculus was the concept of fluents and fluxions. A *fluent* is a moving or flowing quantity; its *fluxion* is its rate of flow, which we now call its velocity or derivative. In his *Fluxions*, written in Latin in 1671 and published in 1742 (an English translation appeared in 1736), he replaced the notation p for velocity by \dot{x} , a notation still used in mechanics and in the calculus of variations. Newton's notation for the opposite operation, finding a fluent from the fluxion has been abandoned: Where we write $\int x(t) dt$, he wrote \dot{x} .

The first problem in the *Fluxions* is: *The relation of the flowing quantities to one another being given, to determine the relation of their fluxions.* The rule given for solving this problem is to arrange the equation that expresses the given relation (assumed algebraic) in powers of one of the variables, say x , multiply its terms by any arithmetic progression (that is, the first power is multiplied by c , the square by $2c$, the cube by $3c$, etc.), and then multiply by \dot{x}/x . After this operation has been performed for each of the variables, the sum of all the resulting terms is set equal to zero.

Newton illustrated this operation with the relation $x^3 - ax^2 + axy - y^2 = 0$, for which the corresponding fluxion relation is $3x^2\dot{x} - 2ax\dot{x} + \dot{a}xy + a\dot{x}y + ax\dot{y} - 2y\dot{y} = 0$, and by numerous examples of finding tangents to well-known curves such as the spiral and the cycloid. Newton also found their curvatures and areas. The combination of these techniques with infinite series was important, since fluents often could not be found in finite terms. For example, Newton found that the area under the curve $\dot{z} = 1/(1+x^2)$ was given by the Jyeshtadeva-Gregory series $z = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \frac{1}{7}x^7 + \dots$.

Later exposition of the calculus. Newton made an attempt to explain fluxions in terms that would be more acceptable logically, calling it the "method of first and last ratios," in his treatise on mechanics, the *Philosophiae naturalis principia mathematica* (*Mathematical Principles of Natural Philosophy*), where he said,

Quantities, and the ratios of quantities, which in any finite time converge continually toward equality, and before the end of that

time approach nearer to each other than by any given difference, become ultimately equal.

If you deny it, suppose them to be ultimately unequal, and let D be their ultimate difference. Therefore they cannot approach nearer to equality than by that given difference D ; which is contrary to the supposition.

If only the phrase *become ultimately equal* had some clear meaning, as Newton seemed to assume, this argument might have been convincing. As it is, it comes close to being a definition of *ultimately equal*, or, as we would say, equal in the limit. Newton came close to stating the modern concept of a limit, when he described the “ultimate ratios” (derivatives) as “limits towards which the ratios of quantities decreasing without limits do always converge, and to which they approach nearer than by any given difference.” Here one can almost see the “arbitrarily small ϵ ” that plays the central role in the concept of a limit.

2.2. Gottfried Wilhelm von Leibniz. Leibniz believed in the reality of infinitesimals, quantities so small that any finite sum of them is still less than any assignable positive number, but which are nevertheless not zero, so that one is allowed to divide by them. The three kinds of numbers (finite, infinite, and infinitesimal) could, in Leibniz’ view, be multiplied by one another, and the result of multiplying an infinite number by an infinitesimal might be any one of the three kinds. This position was rejected in the nineteenth century but was resurrected in the twentieth century and made logically sound. It lies at the heart of what is called *nonstandard analysis*, a subject that has not penetrated the undergraduate curriculum. The radical step that must be taken in order to believe in infinitesimals is a rejection of the Archimedean axiom that for any two positive quantities of the same kind a sufficient number of bisections of one will lead to a quantity smaller than the second. This principle was essential to the use of the method of exhaustion, which was one of the crowning glories of Euclidean geometry. It is no wonder that mathematicians were reluctant to give it up.

Leibniz invented the expression dx to indicate the difference of two infinitely close values of x , dy to indicate the difference of two infinitely close values of y , and dy/dx to indicate the ratio of these two values. This notation was beautifully intuitive and is still the preferred notation for thinking about calculus. Its logical basis at the time was questionable, since it avoided the objections listed above by claiming that the two quantities have not vanished at all but have yet become less than any assigned positive number. However, at the time, consistency would have been counterproductive in mathematics and science.

The integral calculus and the fundamental theorem of calculus flowed very naturally from Leibniz’ approach. Leibniz could argue that the ordinates to the points on a curve represent infinitesimal rectangles of height y and width dx , and hence finding the area under the curve—“summing all the lines in the figure”—amounted to summing infinitesimal differences in area dA , which collapsed to give the total area. Since it was obvious that on the infinitesimal level $dA = ydx$, the fundamental theorem of calculus was an immediate consequence. Leibniz first set it out in geometric form in a paper on quadratures in the 1693 *Acta eruditorum*. There he considered two curves: one, which we denote $y = f(x)$ with its graph above a horizontal axis, the other, which we denote $z = F(x)$, with its graph below

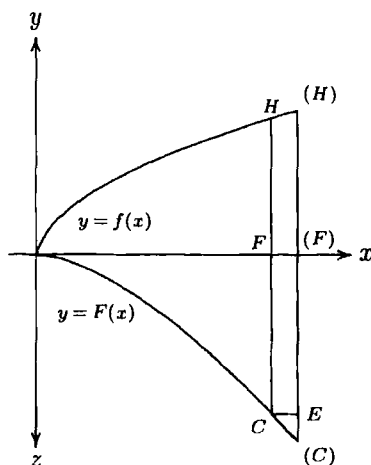


FIGURE 5. Leibniz' proof of the fundamental theorem of calculus.

the horizontal axis.² The second curve has an ordinate proportional to the area under the first curve. That is, for a positive constant a , having the dimension of length, $aF(x)$ is the area under the curve $y = f(x)$ from the origin up to the point with abscissa x . As we would write the relation now,

$$aF(x) = \int_0^x f(t) dt.$$

In this form the relation is dimensionally consistent. What Leibniz proved was that the curve $z = F(x)$, which he called the *quadratrix* (squarer), could be constructed from its infinitesimal elements. In Fig. 5 the parentheses around letters denote points at an infinitesimal distance from the points denoted by the same letters without parentheses. In the infinitesimal triangle $CE(C)$ the line $E(C)$ represents dF , while the infinitesimal quadrilateral $HF(F)(H)$ represents dA , the element of area under the curve. The lines $F(F)$ and CE both represent dx . Leibniz argued that by construction, $a dF = f(x) dx$, and so $dF : dx = f(x) : a$. That meant that the quadratrix could be constructed by antidifferentiation.

Leibniz eventually abbreviated the sum of all the increments in the area (that is, the total area) using an elongated S, so that $A = \int dA = \int y dx$. Nearly all the basic rules of calculus for finding the derivatives of the elementary functions and the derivatives of products, quotients, and so on, were contained in Leibniz' 1684 paper on his method of finding tangents. However, he had certainly obtained these results much earlier. His collected works contain a paper written in Latin with the title *Compendium quadraturae arithmeticae*, to which the editor assigns a date of 1678 or 1679. This paper shows Leibniz' approach through infinitesimal differences and their sums and suggests that it was primarily the problem of squaring the circle and other conic sections that inspired this work. The work consists of 49 propositions and two problems. Most of the propositions are stated without

² The vertical axis is to be assumed positive in both directions from the origin. We are preserving in Fig. 5 only the lines needed to explain Leibniz' argument. He himself merely labeled points on the two curves with letters and referred to those letters.

proof; they contain the basic results on differentiation and integration of elementary functions, including the Taylor series expansions of logarithms, exponentials, and trigonometric functions. Although the language seems slightly archaic, one can easily recognize a core of standard calculus here.

Later reflections on the calculus. Like Newton, Leibniz was forced to answer objections to the new methods of the calculus. In the *Acta eruditorum* of 1695 Leibniz published (in Latin) a "Response to certain objections raised by Herr Bernardo Niewentiit regarding differential or infinitesimal methods." These objections were three: (1) that certain infinitely small quantities were discarded as if they were zero (this principle was set forth as fundamental in the following year in the textbook of calculus by the Marquis de l'Hospital); (2) the method could not be applied when the exponent is a variable; and (3) the higher-order differentials were inconsistent with Leibniz' claim that only geometry could provide a foundation. In answer to the first objection Leibniz attempted to explain different orders of infinitesimals, pointing out that one could neglect all but the lowest orders in a given equation. To answer the second, he used the binomial theorem to demonstrate how to handle the differentials dx , dy , dz when $y^x = z$. To answer the third, Leibniz said that one should not think of $d(dx)$ as a quantity that fails to yield a (finite) quantity even when multiplied by an infinite number. He pointed out that if x varies geometrically when y varies arithmetically, then $dx = (x dy)/a$ and $ddx = (dx dy)/a$, which makes perfectly good sense.

2.3. The disciples of Newton and Leibniz. Newton and Leibniz had disciples who carried on their work. Among Newton's followers was Roger Cotes (1682–1716), who oversaw the publication of a later edition of Newton's *Principia* and defended Newton's inverse square law of gravitation in a preface to that work. He also fleshed out the calculus with some particular results on plane loci and considered the extension of functions defined by power series to complex values, deriving the important formula $i\phi = \log(\cos \phi + i \sin \phi)$, where $i = \sqrt{-1}$. Another of Newton's followers was Brook Taylor (1685–1731), who developed a calculus of finite differences that mirrors in many ways the "continuous" calculus of Newton and Leibniz and is of both theoretical and practical use today. Taylor is famous for the infinite power series representation of functions that now bears his name. It appeared in his 1715 treatise on finite differences. We have already seen, however, that many particular "Taylor series" were known to Newton and Leibniz; Taylor's merit is to have recognized a general way of producing such a series in terms of the derivatives of the generating function. This step, however, was also taken by Johann Bernoulli.

Leibniz also had a group of active and intelligent followers who continued to develop his ideas. The most prominent of these were the Bernoulli brothers Jakob (1654–1705) and Johann (1667–1748), citizens of Switzerland, between whom relations were not always cordial. They investigated problems that arose in connection with calculus and helped to systematize, extend, and popularize the subject. In addition, they pioneered new mathematical subjects such as the calculus of variations, differential equations, and the mathematical theory of probability. A French nobleman, the Marquis de l'Hospital (1661–1704), took lessons from Johann Bernoulli and paid him a salary in return for the right to Bernoulli's mathematical discoveries. As a result, Bernoulli's discovery of a way of assigning values to what are now called indeterminate forms appeared in L'Hospital's 1696 textbook *Analyse*

des infiniment petits (*Infinitesimal Analysis*) and has ever since been known as L'Hospital's rule. Like the followers of Newton, who had to answer the objections of Bishop Berkeley (see Section 3 below) Leibniz' followers encountered objections from Michel Rolle (1652–1719), which were answered by Johann Bernoulli with the claim that Rolle didn't understand the subject.

The priority dispute. One of the better-known and less edifying incidents in the history of mathematics is the dispute between the disciples of Newton and those of Leibniz over the credit for the invention of the calculus. Although Newton had discovered the calculus by the early 1670s and had described it in a paper sent to James Collins, the librarian of the Royal Society, he did not publish his discoveries until 1687. Leibniz made his discoveries a few years later than Newton but published some of them earlier, in 1684. Newton's vanity was wounded in 1695 when he learned that Leibniz was regarded on the Continent as the discoverer of the calculus, even though Leibniz himself made no claim to this honor. In 1699 a Swiss immigrant to England, Nicolas Fatio de Duillier (1664–1753), suggested that Leibniz had seen Newton's paper when he had visited London and talked with Collins in 1673. (Collins died in 1683, before his testimony in the matter was needed.) This unfortunate rumor poisoned relations between Newton and Leibniz and their followers. In 1711–1712 a committee of the Royal Society (of which Newton was President) investigated the matter and reported that it believed Leibniz had seen certain documents that in fact he had not seen. Relations between British and Continental mathematicians reached such a low ebb that Newton deleted certain laudatory references to Leibniz from the third edition of his *Principia*. This dispute confirmed the British in the use of the clumsy Newtonian notation for more than a century, a notation far inferior to Leibniz's elegant and intuitive symbolism. But in the early nineteenth century the impressive advances made by Continental scholars such as Euler, Lagrange, and Laplace won over the British mathematicians, and scholars such as William Wallace (1768–1843) rewrote the theory of fluxions in terms of the theory of limits. Wallace asserted that there was never any need to introduce motion and velocity into this theory, except as illustrations, and that indeed Newton himself used motion only for illustration, recasting his arguments in terms of limits when rigor was needed (see Panteki, 1987, and Craik, 1999). Eventually, even the British began using the term *integral* instead of *fluent* and *derivative* instead of *fluxion*, and these Newtonian terms became mathematically part of a dead language.

Certain relevant facts were concealed by the terms in which the priority dispute was cast. One of these is the extent to which Fermat, Descartes, Cavalieri, Pascal, Roberval, and others had developed the techniques in isolated cases that were to be unified by the calculus as we know it now. In any case, Newton's teacher Isaac Barrow had the insight into the connection between subtangents and area before either Newton or Leibniz thought of it. Barrow's contributions were shunted aside in the heat of the dispute; their significance has been pointed out by Feingold (1993).

Early textbooks on calculus. The secure place of calculus in the mathematical curriculum was established by the publication of a number of excellent textbooks. One of the earliest was the *Analyse des infiniment petits*, mentioned above, which was published by the Marquis de l'Hospital in 1696.

Most students of calculus know the Maclaurin series as a special case of the Taylor series. Its discoverer was a Scottish contemporary of Taylor, Colin Maclaurin (1698–1746), whose treatise on fluxions (1742) contained a thorough and rigorous exposition of calculus. It was written partly as a response to the philosophical attacks on the foundations of calculus by the philosopher George Berkeley.

The Italian textbook *Istituzioni analitiche ad uso della gioventù italiana* (*Analytic Principles for the Use of Italian Youth*) became a standard treatise on analytic geometry and calculus and was translated into English in 1801. Its author was Maria Gaetana Agnesi, who was mentioned in Chapter 4 as one of the first women to achieve prominence in mathematics.

The definitive textbooks of calculus were written by the greatest mathematician of the eighteenth century, the Swiss scholar Leonhard Euler. In his 1748 *Introductio in analysin infinitorum*, a two-volume work, Euler gave a thorough discussion of analytic geometry in two and three dimensions, infinite series (including the use of complex variables in such series), and the foundations of a systematic theory of algebraic functions. The modern presentation of trigonometry was established in this work. The *Introductio* was followed in 1755 by *Institutiones calculi differentialis* and a three-volume *Institutiones calculi integralis* (1768–1774), which included the entire theory of calculus and the elements of differential equations, richly illustrated with challenging examples. It was from Euler's textbooks that many prominent nineteenth-century mathematicians such as the Norwegian genius Niels Henrik Abel first encountered higher mathematics, and the influence of Euler's books can be traced in their work.

The state of the calculus around 1700. Most of what we now know as calculus—rules for differentiating and integrating elementary functions, solving simple differential equations, and expanding functions in power series—was known by the early eighteenth century and was included in the standard textbooks just mentioned. Nevertheless, there was much unfinished work. We list here a few of the open questions:

Nonelementary integrals. Differentiation of elementary functions is an algorithmic procedure, and the derivative of any elementary function whatsoever, no matter how complicated, can be found if the investigator has sufficient patience. Such is not the case for the inverse operation of integration. Many important elementary functions, such as $(\sin x)/x$ and e^{-x^2} , are not the derivatives of elementary functions. Since such integrals turned up in the analysis of some fairly simple motions, such as that of a pendulum, the problem of these integrals became pressing.

Differential equations. Although integration had originally been associated with problems of area and volume, because of the importance of differential equations in mechanical problems the solution of differential equations soon became the major application of integration. The general procedure was to convert an equation to a form in which the derivatives could be eliminated by integrating both sides (reduction to quadratures). As these applications became more extensive, more and more cases began to arise in which the natural physical model led to equations that could not be reduced to quadratures. The subject of differential equations began to take on a life of its own, independent of the calculus.

Foundational difficulties. The philosophical difficulties connected with the use of infinitesimal methods were paralleled by mathematical difficulties connected with the application of the algebra of finite polynomials to infinite series. These difficulties

were hidden for some time, and for a blissful century mathematicians and physicists operated formally on power series as if they were finite polynomials. They did so even though it had been known since the time of Oresme that the partial sums of the harmonic series $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ grow arbitrarily large.

3. Branches and roots of the calculus

The calculus grew organically, sending forth branches while simultaneously putting down firm roots. The roots were the subject of philosophical speculation that eventually led to new mathematics as well, but the branches were natural outgrowths of pure mathematics that appeared very early in the history of the subject. We begin this section with the branches and will end it with the roots.

3.1. Ordinary differential equations. Ordinary differential equations arose almost as soon as there was a language (differential calculus) in which they could be expressed.³ These equations were used to formulate problems from geometry and physics in the late seventeenth century, and the natural approach to solving them was to apply the integral calculus, that is, to reduce a given equation to quadratures. Leibniz, in particular, developed the technique now known as separation of variables as early as 1690 (Grosholz, 1987). In the simplest case, that of an ordinary differential equation of first order and first degree, one is seeking an equation $f(x, y) = c$, which may be interpreted as a conservation law if x and y are functions of time having physical significance. The conservation law is expressed as the differential equation

$$\frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy = 0.$$

The resulting equation is known as an exact differential equation. To solve this equation, one has only to integrate the first differential with respect to x , adding an arbitrary function $g(y)$ to the solution, then differentiate with respect to y and compare the result with $\frac{\partial f}{\partial y}$ in order to get an equation for $g'(y)$, which can then be integrated.

If all equations were this simple, differential equations would be a very trivial subject. Unfortunately, it seems that nature tries to confuse us, multiplying these equations by arbitrary functions $\mu(x, y)$. That is, when an equation is written down as a particular case of a physical law, it often looks like

$$M(x, y) dx + N(x, y) dy = 0,$$

where $M(x, y) = \mu(x, y) \frac{\partial f}{\partial x}$ and $N(x, y) = \mu(x, y) \frac{\partial f}{\partial y}$, and no one can tell from looking at M just which factors in it constitute μ and which constitute $\frac{\partial f}{\partial x}$. To take the simplest possible example, the mass y of a radioactive substance that remains undecayed in a sample after time t satisfies the equation

$$dy - ky dx = 0,$$

where k is a constant. The mathematician's job is to get rid of $\mu(x, y)$ by looking for an "integrating factor" that will make the equation exact. One integrating factor for this equation is $1/y$; another is e^{-kx} .

³ This subsection is a summary of an unpublished paper that can be found in full at the following website: <http://www.emba.uvm.edu/~cooke/ckthm.pdf>

It appeared at a very early stage that finding an integrating factor is not in general possible, and both Newton and Leibniz were led to the use of infinite series with undetermined coefficients to solve such equations. Later, Maclaurin, was to warn against too hasty recourse to infinite series, saying that certain integrals could be better expressed geometrically as the arc lengths of various curves. But the idea of replacing a differential equation by a system of algebraic equations was very attractive. The earliest examples of series solutions were cited by Feigenbaum (1994). In his *Fluxions*, which was written in 1671 and left unpublished during his lifetime (see Whiteside, 1967, Vol. 3, p. 99), Newton considered the linear differential equation that we would now write as

$$\frac{dy}{dx} = 1 - 3x + x^2 + (1 + x)y.$$

Newton wrote it as $n/m = 1 - 3x + y + xx + xy$ and found that

$$y = x - x^2 + \frac{1}{3}x^3 - \frac{1}{6}x^4 + \frac{1}{30}x^5 - \frac{1}{45}x^6 - \dots$$

Similarly, in a paper published in the *Acta eruditorum* in 1693 (Gerhardt, 1971, Vol. 5, p. 287), Leibniz studied the differential equations for the logarithm and the arcsine in order to obtain what we now call the Maclaurin series of the logarithm, exponential, and sine functions. For example, he considered the equation $a^2 dy^2 = a^2 dx^2 + x^2 dy^2$ and assumed that $x = by + cy^3 + ey^5 + fy^7 + \dots$, thereby obtaining the series that represents the function $x = a \sin(y/a)$. Neither Newton nor Leibniz mentioned that the coefficients in these series were the derivatives of the functions represented by the series divided by the corresponding factorials. However, that realization came to Johann Bernoulli very soon after the publication of Leibniz' work. In a letter to Leibniz dated September 2, 1694 (Gerhardt, 1971, Vol. 3/1, p. 350), Bernoulli described essentially what we now call the Taylor series of a function. In the course of this description, he gave in passing what became a standard definition of a function, saying, "I take n to be a quantity formed in an arbitrary manner from variables and constants." Leibniz had used this word as early as 1673, and in an article in the 1694 *Acta eruditorum* had defined a function to be "the portion of a line cut off by lines drawn using only a fixed point and a given point lying on a curved line." As Leibniz said, a given curve defines a number of functions: its abscissas, its ordinates, its subtangents, and so on. The problem that differential equations solve is to reconstruct the curve given the ratio between two of these functions.

In classical terms, the solution of a differential equation is a function or family of functions. Given that fact, the ways in which a function can be presented become an important issue. With the modern definition of a function and the familiar notation, one might easily forget that in order to apply the theory of functions it is necessary to deal with particular functions, and these must be *presented* somehow. Bernoulli's description addresses that issue, although it leaves open the question of what methods of combining variables and constants are legal.

A digression on time. The Taylor series of a given function can be generated knowing the values of the function over any interval of the independent variable, no matter how short. Thus, a quantity represented by such a series is determined for all values of the independent variable when the values are given on any interval

at all. Given that the independent variable is usually time, that property corresponds to physical determinacy: Knowing the full state of a physical quantity for some interval of time determines its values for all time. Lagrange, in particular, was a proponent of power series, for which he invented the term *analytic function*. However, as we now know, the natural domain of analytic function theory is the complex numbers. Now in mechanics the independent variable represents time, and that fact raises an interesting question: Why should time be a complex variable? How do complex numbers turn out to be relevant to a problem where only real values of the variables have any physical meaning? To this question the eighteenth- and nineteenth-century mathematicians gave no answer. Indeed, it does not appear that they even asked the question very often. Extensive searches of the nineteenth-century literature by the present author have produced only the following comments on this interesting question, made by Weierstrass in 1885 (see his *Werke*, Bd. 3, S. 24):

It is very remarkable that in a problem of mathematical physics where one seeks an unknown function of two variables that, in terms of their physical meaning, can have only real values and is such that for a particular value of one of the variables the function must equal a prescribed function of the other, an expression often results that is an analytic function of the variable and hence also has a meaning for complex values of the latter.

It is indeed very remarkable, but neither Weierstrass nor anyone since seems to have explained the mystery. But, just as complex numbers were needed to produce the three real roots of a cubic equation, it may not have seemed strange that the complex-variable properties of solutions of differential equations are relevant, even in the study of problems generated by physical considerations involving only real variables. Time is, however, sometimes represented as a two-dimensional quantity, in connection with what are known as Gibbs random fields.

3.2. Partial differential equations. In the middle of the eighteenth century mathematical physicists began to consider problems involving more than one independent variable. The most famous of these is the vibrating string problem discussed by Euler, d'Alembert, and Daniel Bernoulli (1700–1782, son of Johann Bernoulli) during the 1740s and 1750s. This problem led to the one-dimensional wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},$$

with the initial condition $u(x, 0) = f(x)$, $\frac{\partial u}{\partial t}(x, 0) = 0$, which Bernoulli solved in the form of an infinite trigonometric series

$$\sum_{n=1}^{\infty} a_n \sin nx \cos nct,$$

the a_n being chosen so that $\sum_{n=1}^{\infty} a_n \sin nx = f(x)$.⁴

With this problem, partial differential equations arose, leading to new methods of solution. The developments that grew out of trigonometric-series techniques are

⁴ This solution was criticized by Euler, leading to a debate over the allowable methods of defining functions and the proper definition of a function.

discussed in Chapter 17, along with the development of real analysis in general. For the rest of the present section, we confine our discussion to power-series techniques.

The heat equation

$$\frac{\partial u}{\partial t} = a \frac{\partial^2 u}{\partial x^2}$$

was the first partial differential equation to which the power-series method was applied. Fourier used this method to produce the solution

$$u(x, t) = \sum_{r=0}^{\infty} \frac{\varphi^{(2r)}(x)}{r!} t^r$$

when $a = 1$, without realizing that this solution “usually” diverges.

It was not until the nineteenth century that mathematicians began to worry about the convergence of their series solutions. Then Cauchy and Weierstrass produced proofs that the series do converge for ordinary differential equations, provided that the coefficients have convergent series representations. For partial differential equations, it turned out that the form of the equation had some influence. Weierstrass' student Sof'ya Kovalevskaya showed that in general the power series solution for the heat equation diverges if the initial temperature distribution is prescribed, even when that temperature is an analytic function of position. She showed, however, that the series converges if the temperature and temperature gradient at one point are prescribed as analytic functions of time. More generally, she showed that the power-series solution of any equation in “normal form” (solvable for a pure derivative of order equal to the order of the equation) would converge.

3.3. Calculus of variations. The notion of function lies at the heart of calculus. The usual picture of a function is of one *point* being mapped to another *point*. However, the independent variable in a function can be a curve or surface as well as a point. For example, given a curve γ that is the graph of a function $y = f(x)$ between $x = a$ and $x = b$, we can define its length as

$$\Lambda(\gamma) = \int_a^b \sqrt{1 + (f'(x))^2} \, dx.$$

One of the important problems in the history of geometry has been to pick out the curve γ that minimizes $\Lambda(\gamma)$ and satisfies certain extra conditions, such as joining two fixed points P and Q on a surface or enclosing a fixed area A . The calculus technique of “setting the derivative equal to zero” needs to be generalized for such problems, and the techniques for doing so constitute the calculus of variations. The history of this outgrowth of the calculus has been studied very thoroughly in a number of classic works, such as Woodhouse (1810),⁵ Todhunter (1861), and Goldstine (1980), as well as many articles, such as Kreyszig (1993).

⁵ The treatise of Woodhouse is a textbook as much as a history, and its last chapter is a set of 29 examples posed as exercises for the reader with solutions provided. The book also marks an important transition in British mathematics. Woodhouse says in the preface that, “In a former Work, I adopted the foreign notation...”. The foreign notation was the Leibniz notation for differentials, in preference to the dot above the letter that Newton used to denote his fluxions. He says that he found this notation even more necessary in calculus of variations, since he would otherwise have had to adopt some new symbol for Lagrange's variation. But he then goes on to marvel that Lagrange had taken the reverse step of introducing Newton's fluxion notation into the calculus of variations.

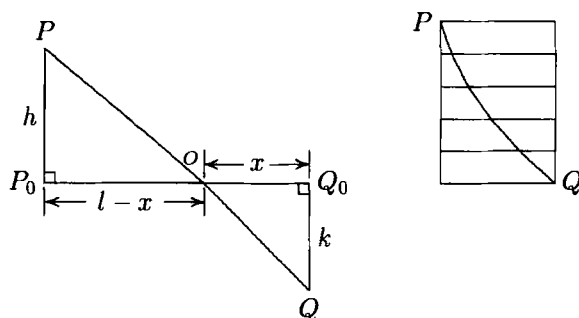


FIGURE 6. Left: Fermat's principle. Choosing the point O so that the time of travel from P to Q through O is a minimum. Right: the principle applied layer by layer when the speed increases proportionally to the square root of the distance of descent.

As with the ordinary calculus, the development of calculus of variations proceeded from particular problems solved by special devices to general techniques and algorithms based on theoretical analysis and rigorous proof. In the seventeenth century there were three such special problems that had important consequences. The first was the brachystochrone (shortest-time) problem for an object crossing an interface between two media while moving from one point to another. In the simplest case (Fig. 6), the interface is a straight line, and the point O is to be chosen so that the time required to travel from P to O at speed v , then from O to Q at speed w , is minimized. If the two speeds are not the same, it is clear that the path of minimum time will not be a straight line, since time can be saved by traveling a slightly longer distance in the medium in which the speed is greater.

The second problem, that of finding the cross-sectional shape of the optimally streamlined body moving through a resisting medium, is discussed in the scholium to Proposition 34 (Theorem 28) of Book 2 of Newton's *Principia*.

Fermat's principle, which asserts that the path of a light ray is the one that requires least time, came into play in a challenge problem stated by Johann Bernoulli in 1696: To find the curve down which a frictionless particle will slide from point P to point Q under the influence of gravity in minimal time. Since the speed of a falling body is proportional to the square root of the distance fallen, Bernoulli reasoned that the sine of the angle between the tangent and the vertical would be proportional to the square root of the vertical coordinate (assuming the vertical axis directed downward). (Recall that ibn Sahl, al-Haytham, Harriot, Snell, and Descartes had all derived the law of refraction which asserts that the ratio of the sines of the angles of incidence and refraction at an interface are proportional to the velocities in the two media.) In that way he arrived at a differential equation for the curve:

$$\frac{dy}{dx} = \sqrt{\frac{y}{a+y}}.$$

(We have taken y as the vertical coordinate. Bernoulli apparently took x .) He recognized this equation as the differential equation of a cycloid and thus came to the fascinating conclusion that this curve, which Huygens had studied because it enabled a clock to keep theoretically perfect time (the tautochrone property), also had the brachystochrone property. The challenge problem was solved by Bernoulli

himself, his brother Jakob, and by both Newton and Leibniz.⁶ According to Woodhouse (1810, p. 150), Newton's anonymously submitted solution was so concise and elegant that Johann Bernoulli knew immediately who it must be from. He wrote, "Even though the author, from excessive modesty, does not give his name, we can nevertheless tell certainly by a number of signs that it is the famous Newton; and even if these signs were not present, seeing a small sample would suffice to recognize him, as *ex ungue Leonem*."⁷

Euler. Variational problems were categorized and systematized by Euler in a large treatise in 1744 named *Methodus inveniendi lineas curvas (A Method of Finding Curves)*. In this treatise Euler set forth a series of problems of increasing complexity, each involving the finding of a curve having certain extremal properties, such as minimal length among all curves joining two points on a given surface.⁸ Proposition 3 in Chapter 2, for example, asks for the minimum value of an integral $\int Z dx$, where Z is a function of variables, x , y , and $p = y' = \frac{dy}{dx}$. Based on his previous examples, Euler derived the differential equation

$$0 = N dx - dP,$$

where $dZ = M dx + N dy + P dp$ is the differential of the integrand Z . Since $N = \frac{\partial Z}{\partial y}$ and $P = \frac{\partial Z}{\partial p}$, this equation could be written in the form that is now the basic equation of the calculus of variations, and is known as Euler's equation:

$$\frac{\partial Z}{\partial y} = \frac{d}{dx} \left(\frac{\partial Z}{\partial y'} \right).$$

In Chapter 3, Euler generalized this result by allowing Z to depend on additional parameters and applied his result to find minimal surfaces. In an appendix he studied elastic curves and surfaces, including the problem of the vibrating membrane. This work was being done at the very time when Euler's friend Daniel Bernoulli was studying the simpler problem of the vibrating string. In a second appendix he showed how to derive the equations of mechanics from variational principles, thus providing a unifying mathematical principle that applied to both optics (Fermat's principle) and mechanics.

Lagrange. The calculus of variations acquired "variations" and its name as the result of a letter written by Lagrange to Euler in 1755. In that letter, Lagrange generalized Leibniz' differentials from points to curves, using the Greek δ instead of the Latin d to denote them. Thus, if $y = f(x)$ was a curve, its *variation* δy was a small perturbation of it. Just as dy was a small change in the value of y at a point, δy was a small change in all the values of y at all points. The variation operator δ can be manipulated quite easily, since it commutes with differentiation and integration: $\delta y' = (\delta y)'$ and $\delta \int Z dx = \int \delta Z dx$. With this operator, Euler's equation and its many applications, were easy to derive. Euler immediately recognized the usefulness of what Lagrange had done and gave the new theory the name it has borne ever since: calculus of variations.

Lagrange also considered extremal problems with constraint and introduced the famous Lagrange multipliers as a way of turning these relative (constrained)

⁶ Newton apparently recognized structural similarities between this problem and his own optimal-streamlining problem (see Goldstine, 1980, pp. 7-35).

⁷ A Latin proverb much in vogue at the time. It means literally "from [just] the claw [one can recognize] the Lion."

⁸ This problem was Example 4 in Chapter 4 of the treatise.

extrema into absolute (unconstrained) extrema. Euler had given an explanation of this process earlier. Woodhouse (1810, p. 79) thought that Lagrange's systematization actually deprived Euler's ideas of their simplicity.

Second-variation tests for maxima and minima. Like the equation $f'(x) = 0$ in calculus, the Euler equation is only a necessary condition for an extremal, not sufficient, and it does not distinguish between maximum, minimum, and neither. In general, however, if Euler's equation has only one solution, and there is good reason to believe that a maximum or minimum exists, the solution of the Euler equation provides a basis to proceed in practice. Still, mathematicians were bound to explore the question of distinguishing maxima from minima. Such investigations were undertaken by Lagrange and Legendre in the late eighteenth century.

In 1786 Legendre was able to show that a sufficient condition for a minimum of the integral

$$I(y) = \int_a^b f(x, y, y') dx,$$

at a function satisfying Euler's necessary condition, was $\frac{\partial^2 f}{\partial y'^2} > 0$ for all x and that a sufficient condition for a maximum was $\frac{\partial^2 f}{\partial y'^2} < 0$.

In 1797 Lagrange published a comprehensive treatise on the calculus, in which he objected to some of Legendre's reasoning, noting that it assumed that certain functions remained finite on the interval of integration (Dorofeeva, 1998, p. 209).⁹

Jacobi: sufficiency criteria. The second-variation test is strong enough to show that a solution of the Euler equation really is an extremal among the smooth functions that are "nearby" in the sense that their values are close to those of the solution and their derivatives also take values close to those of the derivative of the solution. Such an extremal was called a *weak extremal* by Adolf Kneser (1862–1930). Jacobi had the idea of replacing the curve $y(x)$ that satisfied Euler's equation with a family of such curves depending on parameters (two in the case we have been considering) $y(x, \alpha_1, \alpha_2)$ and replacing the nearby curves $y + \delta y$ and $y' + \delta y'$ with values corresponding to different parameters. In 1837—see Dorofeeva (1998) or Fraser (1993)—he finally solved the problem of finding sufficient conditions for an extremal. He included his solution in the lectures on dynamics that he gave in 1842, which were published in 1866, after his death. The complication that had held up Jacobi and others was the fact that sometimes the extremals with given endpoints are not unique. The most obvious example is the case of great circles on the sphere, which satisfy the Euler equations for the integral that gives arc length subject to fixed endpoints. If the endpoints happen to be antipodal points, all great circles passing through the two points have the same length. Weierstrass was later to call such pairs of points *conjugate points*. Jacobi gave a differential equation whose solutions had zeros at these points and showed that Legendre's criterion was correct, provided that the interval $(a, b]$ contained no points conjugate to a .

Weierstrass and his school. A number of important advances in the calculus of variations were due to Karl Weierstrass, such as the elimination of some of the more

⁹ More than that was wrong, however, since great circles on a sphere satisfy Legendre's criteria, but do not give a minimum distance between their endpoints if they are more than 180° long.

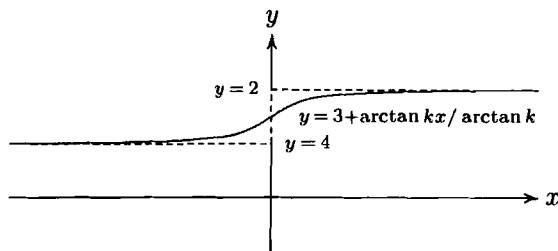


FIGURE 7. The functional $\Phi(y) = \int_{-1}^{+1} (xy'(x))^2 dx$ does not assume its minimum value for continuously differentiable functions $y(x)$ satisfying $y(-1) = 2$, $y(+1) = 4$. The limiting position of a minimizing sequence is the dashed line.

restrictive assumptions about differentiability and taking account of the distinction between a lower bound and a minimum.¹⁰

An important example in this connection was Riemann's use of *Dirichlet's principle* to prove the Riemann mapping theorem, which asserts that any simply connected region in the plane except the plane itself can be mapped conformally onto the unit disk $\Delta = \{(x, y) : x^2 + y^2 < 1\}$. That principle required the existence of a real-valued function $u(x, y)$ that minimizes the integral

$$\iint_{\Delta} \left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 dx dy$$

among all functions $u(x, y)$ taking prescribed values on the boundary of the disk. That function is the unique harmonic function in Δ with the given boundary values. In 1870 Weierstrass called attention to the integral

$$\Phi(\varphi) = \int_{-1}^{+1} x^2 (\varphi'(x))^2 dx,$$

which when combined with the boundary condition $\varphi(-1) = a$, $\varphi(+1) = b$, can be made arbitrarily small by the function

$$\varphi(x) = \frac{a+b}{2} + \frac{b-a}{2} \frac{\arctan(kx)}{\arctan(k)},$$

yet (if $a \neq b$) cannot be zero for any function φ satisfying the boundary conditions and such that φ' exists at every point.

Weierstrass' example was a case where it was necessary to look outside the class of smooth functions for a minimum of the functional. The limiting position of the graphs of the functions for which the integral approximates its minimum value consists of the two horizontal lines from $(-1, a)$ to $(0, a)$, from $(0, b)$ to $(+1, b)$, and the section of the y -axis joining them (see Fig. 7).

Weierstrass thought of the smoothness assumptions as necessary evils. He recognized that they limited the generality of the results, yet he saw that without them no application of the calculus was possible. The result is a certain vagueness about the formulation of minimal principles in physics. A certain functional must be

¹⁰ This distinction was pointed out by Gauss as early as 1799, in his criticism of d'Alembert's 1746 proof of the fundamental theorem of algebra.

a minimum *assuming* that all the relevant quantities are differentiable a sufficient number of times. Obviously, if a functional can be extended to a wider class of functions in a natural way, the minimum reached may be smaller, or the maximum larger. To make the restrictions as weak as possible, Weierstrass imposed the condition that the partial derivatives of the integrand should be continuous at corners. An extremal among all functions satisfying these less restrictive hypotheses was called a *strong* extremal. The corner condition was also found by G. Erdmann, a teacher at the Gymnasium in Königsberg, who proved that Jacobi's sufficient condition for a weak extremal was also necessary.

3.4. Foundations of the calculus. The British and Continental mathematicians both found the power of the calculus so attractive that they applied and developed it (sending forth new branches), all the while struggling to be clear about the principles they were using (extending its roots). The branches grew more or less continuously from the beginning. The development of the roots was slower and more sporadic. A satisfactory consensus was achieved only late in the nineteenth century, with the full development of real analysis, which is discussed in the Chapter 17.

The source of all the difficulty was the introduction of the infinite into analysis, in the form of infinitesimal reasoning. Leibniz believed in actual infinitesimals, levels of magnitude that were real, not zero, but so small that no accumulation of them could ever exceed any finite quantity. His dx was such an infinitesimal, and a product of two, such as $dx\,dy$ or dx^2 , was a higher-order infinitesimal, so small that no accumulation of such could ever exceed any infinitesimal of the first order. On this view, even though theorems established using calculus were not absolutely accurate, the errors were below the threshold of human perception and therefore could not matter in practice. Newton was probably alluding to Leibniz when in his discussion of the quadrature of curves, he wrote, "Errores quam minimi in rebus mathematicis non sunt contemnendi" ("Errors, no matter how small, are not to be considered in mathematics"). Newton knew that his arguments could have been phrased using the Eudoxan method of exhaustion. In his *Principia* he wrote that he used his method of first and last ratios "to avoid the tediousness of deducing involved demonstrations *ad absurdum*, according to the method of the ancient geometers."

There seemed to be three approaches that would allow the operation that we now know as integration to be performed by antidifferentiation of tangents. One is the infinitesimal approach of Leibniz, characterized by Mancosu (1989) as "static." That is, a tangent is a state or position of a line, namely that of passing through two infinitely near points. The second is Newton's "dynamic" approach, in which a fluxion is the velocity of a moving object. The third is the ancient method of exhaustion. In principle, a reduction of calculus to the Eudoxan theory of proportion is possible. Psychologically, it would involve not only a great deal of tedium, as Newton noted, but also a great deal of unnecessary confusion, which he did not point out. If mathematicians had been shackled by the requirements of this kind of rigor, the amount of geometry and analysis created would have been incomparably less than it was. Still, Newton felt the objection and tried to phrase his exposition of the method of first and last ratios in such a way as not to outrage anyone's logical scruples. He said:

Perhaps it may be objected, that there is no ultimate proportion of evanescent quantities; because the proportion, before the quantities have vanished, is not the ultimate; and when they are vanished, is [not defined]. But by the same argument it may be alleged that a body arriving at a certain place, and there stopping, has no ultimate velocity, because the velocity before the body comes to the place, is not its ultimate velocity; when it has arrived, there is none. But the answer is easy; for by the ultimate velocity is meant that with which the body is moved, neither before it arrives at its last place and the motion ceases, nor after, but at the very instant it arrives.

Was this explanation adequate? Do human beings in fact have any conception of what is meant by an instant of time? Do we have a clear idea of the velocity of a body *at the very instant* when it stops moving? Or do some people only imagine that we do? We are here very close to the arrow paradox of Zeno. At any given instant, the arrow does not move; therefore it is at rest. How can there be a motion (a traversal of a positive distance) as a result of an accumulation of states of rest, in each of which no distance is traveled? Newton's "by the same argument" practically invited the further objection that his attempted explanation merely stated the same fallacy in a new way.

That objection was raised in 1734 by the philosopher George Berkeley¹¹ (1685–1753, Anglican Bishop of Cloyne, Ireland), for whom the city of Berkeley¹² in California is named. Berkeley first took on Newton's fluxions:

It is said that the minutest errors are not to be neglected in mathematics: that the fluxions are celerities [speeds], not proportional to the finite increments, though ever so small; but only to the moments or nascent increments, whereof the proportion alone, and not the magnitude, is considered. And of the aforesaid fluxions there be other fluxions, which fluxions of fluxions are called second fluxions. And the fluxions of the second fluxions are called third fluxions: and so on, fourth, fifth, sixth, &c. *ad infinitum*. Now, as our sense is strained and puzzled with the perception of objects extremely minute, even so the imagination, which faculty derives from sense, is very much strained and puzzled to frame clear ideas of the least particles of time. . . and much more so to comprehend. . . those increments of the flowing quantities. . . in their very first origin, or beginning to exist, before they become finite particles. . . The incipient celerity of an incipient celerity, the nascent augment of a nascent augment, *i.e.*, of a thing which hath no magnitude: take it in what light you please, the clear conception of it will, if I mistake not, be found impossible.

He then proceeded to attack the views of Leibniz:

¹¹ Pronounced "Barkley."

¹² Pronounced "Birkley".

The foreign mathematicians are supposed by some, even of our own, to proceed in a manner less accurate, perhaps, and geometrical, yet more intelligible. . . Now to conceive a quantity infinitely small, that is, infinitely less than any sensible or imaginable quantity or than any the least finite magnitude is, I confess, above my capacity. But to conceive a part of such infinitely small quantity that shall be still infinitely less than it, and consequently though multiplied infinitely shall never equal the minutest finite quantity, is, I suspect, an infinite difficulty to any man whatsoever.

Berkeley analyzed a curve whose area up to x was x^3 (he wrote xxx). If $z - x$ was the increment of the abscissa and $z^3 - x^3$ the increment of area, the quotient would be $z^2 + zx + x^2$. He said that, if $z = x$, of course this last expression is $3x^2$, and that must be the ordinate of the curve in question. That is, its equation must be $y = 3x^2$. But, he pointed out,

[H]erein is a direct fallacy: for, in the first place, it is supposed that the abscissas z and x are unequal, without which supposition no one step could have been made [that is, the division by $z - x$ would have been undefined]; which is a manifest inconsistency, and amounts to the same thing that hath been before considered. . . The great author of the method of fluxions felt this difficulty, and therefore he gave in to those nice abstractions and geometrical metaphysics without which he saw nothing could be done on the received principles. . . It must, indeed, be acknowledged that he used fluxions, like the scaffold of a building, as things to be laid aside or got rid of as soon as finite lines were found proportional to them. . . And what are these fluxions? The velocities of evanescent increments? And what are these same evanescent increments? They are neither finite quantities, nor quantities infinitely small, nor yet nothing. May we not call them the ghosts of departed quantities?

The debate on the Continent. Calculus disturbed the metaphysical assumptions of philosophers and mathematicians on the Continent as well as in Britain. L'Hospital's textbook had made two explicit assumptions: first, that if a quantity is increased or diminished by a quantity that is infinitesimal in comparison with itself, it may be regarded as remaining unchanged. Second, that a curve may be regarded as an infinite succession of straight lines. L'Hospital's justification for these claims was not commensurate with the strength of the assumptions. He merely said:

[T]hey seem so obvious to me that I do not believe they could leave any doubt in the mind of attentive readers. And I could even have proved them easily after the manner of the Ancients, if I had not resolved to treat only briefly things that are already known, concentrating on those that are new. [Quoted by Mancosu, 1989, p. 228]

The idea that $x + dx = x$, implicit in l'Hospital's first assumption, leads algebraically to the equation $dx = 0$ if equations are to retain their previous meaning. Rolle raised this objection and was answered by the claim that dx represents the

distance traveled in an instant of time by an object moving with finite velocity. This debate was carried on in private in the Paris Academy during the first decade of the eighteenth century, and members were at first instructed not to discuss it in public, as if it were a criminal case! Rolle's criticism could be answered, but it was *not* answered at the time. According to Mancosu (1989), the matter was settled in a most unacademic manner, by making l'Hospital into an icon after his death in 1704. His eulogy by Bernard Lebouyer de Fontenelle (1657–1757) simply declared the anti-infinitesimalists wrong, as if the Academy could decide metaphysical questions by fiat, just as it can define what is proper usage in French:

[T]hose who knew nothing of the mysteries of this new infinitesimal geometry were shocked to hear that there are infinities of infinities, and some infinities larger or smaller than others; for they saw only the top of the building without knowing its foundation. [Quoted by Mancosu, 1989, 241]

In the eighteenth century, however, better expositions of the calculus were produced by d'Alembert. In his article on the differential for the famous *Encyclopédie* he wrote that $0/0$ could be equal to anything, and that the derivative $\frac{dy}{dx}$ was not actually 0 divided by 0, but the limit of finite quotients as numerator and denominator tended to zero.

Lagrange's algebraic analysis. The attempt to be clear about infinitesimals or to banish them entirely took many forms during the eighteenth and nineteenth centuries. One of the most prominent (see Fraser, 1987) was Lagrange's exposition of analytic functions. Lagrange understood the term *function* to mean a formula composed of symbols representing variables and arithmetic operations. He argued that "in general" (with certain obvious exceptions) every function $f(x)$ could be expanded as a power series, based on Taylor's theorem, for which he provided his own form of the remainder term. Using an argument that resembles the one given by Ruffini and Abel to prove the insolvability of the quintic, he claimed that the hypothetical expansion

$$\sqrt{x+h} = \sqrt{x} + ph + qh^2 + \cdots + h^{m/n}$$

could not occur, since the left-hand side has only two values, while the right-hand side has n values. In this way, he ruled out fractional exponents. Negative exponents were ruled out by the mere fact that the function was defined at $h = 0$. The determinacy property of analytic functions was used implicitly by Lagrange when he assumed that any zero of a function must have finite order, as we would say (Fraser, 1987, p. 42).

The advantage of confining attention to functions defined by power series is that the derivative and integral of such a function have a perfectly definite meaning. Lagrange advocated it on the grounds that it pointed up the qualitative difference between the new functions produced by infinitesimal analysis: dx was a completely different function from x .

Cauchy's calculus. The modern presentation of calculus owes a great deal to the textbooks of Cauchy, written for his lectures at the École Polytechnique during the 1820s.¹³ Cauchy recognized that calculus could not get by without something

¹³ Although we have mentioned particular results of Cauchy in connection with the solution of algebraic and differential equations, his treatises on analysis are the contributions for which he is

equivalent to infinitesimals. He defined a function $f(x)$ to be continuous if the absolute value of the difference $f(x + \alpha) - f(x)$ “decreases without limit along with that of α .” He continues:

In other words, the function $f(x)$ remains continuous with respect to x in a given interval, if an infinitesimal increase in the variable within this interval always produces an infinitesimal increase in the function itself.

Certain distinctions that we now make to clarify whether x is a fixed point or the increase is thought of as occurring at all points simultaneously are not stated here. In particular, uniform convergence and continuity are assumed but not stated. Cauchy defined a limit in terms of the “successive values attributed to a variable,” approaching a fixed value and ultimately differing from it by an arbitrarily small amount. This definition can be regarded as an informal version of what we now state precisely with deltas and epsilons, and Cauchy is generally regarded, along with Weierstrass, as being one of the people who finally made the foundations of calculus secure. Yet Cauchy’s language clearly presumes that infinitesimals are real. As Laugwitz (1987, p. 272) says:

All attempts to understand Cauchy from a ‘rigorous’ theory of real numbers and functions including uniformity concepts have failed... One advantage of modern theories like the Nonstandard Analysis of Robinson... [which includes infinitesimals] is that they provide consistent reconstructions of Cauchy’s concepts and results in a language which sounds very much like Cauchy’s.

The secure foundation of modern analysis owes much to Cauchy’s treatises. As Grabiner (1981) says, he applied ancient Greek rigor and modern algebraic techniques to derive results from analysis. The contributions of other nineteenth-century mathematicians to this rigor are discussed in Chapter 17.

Questions and problems

16.1. Show that the Madhava–Jyeshtadeva formula given at the beginning of the chapter is equivalent to

$$\theta = \sum_{k=0}^{\infty} (-1)^k \frac{\tan^{2k+1} \theta}{2k+1},$$

or, letting $x = \tan \theta$,

$$\arctan x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{2k+1}.$$

16.2. Consider an ellipse with semiaxes a and b and a circle of radius b , both circle and ellipse lying between a pair of parallel lines a distance $2b$ apart. For every line between the two lines and parallel to them, show that the portion inside the ellipse will be a/b times the portion inside the circle. Use this fact and Cavalieri’s principle to compute the area of the ellipse. This result was given by Kepler.

best remembered. Incidentally, he became a mathematician only after practicing as an engineer for several years.

16.3. Show that the point at which the tangent to the curve $y = f(x)$ intersects the y axis is $y = f(x) - xf'(x)$, and verify that the area under this curve—more precisely, the integral of $f(x) - xf'(x)$ from $x = 0$ to $x = a$ —is twice the area between the curve $y = f(x)$ and the line $ay = f(a)x$ between the points $(0, 0)$ and $(a, f(a))$. This result was used by Leibniz to illustrate the power of his infinitesimal methods.

16.4. Recall that Eudoxus solved the problem of incommensurables by changing the definition of proportion, or rather, *making* a definition to cover cases where no definition existed before. Newton's "theorem" asserting that quantities that approach each other continually (we would say monotonically) and become arbitrarily close to each other in a finite time must become equal in an infinite time assumes that one has a definition of equality at infinity. What is the definition of equality at infinity? Since we cannot *actually* reach infinity, the definition will have to be stated as a potential infinity, that is, a statement about all possible finite times. Formulate the definition and compare Newton's solution of this difficulty with Eudoxus' solution of the problem of incommensurables.

16.5. Draw a square and one of its diagonals. Then draw a very fine "staircase" by connecting short horizontal and vertical line segments in alternation, each segment crossing the diagonal. The total length of the horizontal segments is the same as the side of the square, and the same is true of the vertical segments. Now in a certain intuitive sense these segments approximate the diagonal of the square, seeming to imply that the diagonal of a square equals twice its side, which is absurd. Does this argument show that the method of indivisibles is wrong?

16.6. In the passage quoted from the *Analyst*, Berkeley asserts that the experience of the senses provides the only foundation for our imagination. From that premise he concludes that we can have no understanding of infinitesimals. Analyze whether the premise is true, and if so, whether it implies the conclusion. Assuming that our thinking processes have been shaped by the evolution of the brain, for example, is it possible that some of our spatial and counting intuition is "hard-wired" and not the result of any previous sense impressions? The philosopher Immanuel Kant (1724–1804) thought so. Do we have the power to make correct judgments about spaces and times on scales that we have not experienced? What would Berkeley have said if he had heard Riemann's argument that space may be finite, yet unbounded? How would he have explained the modern computer chip, on which unimaginable amounts of data can be recorded in space far too small for the senses to perceive? Go a step further and consider how quantum mechanics is understood and interpreted.

CHAPTER 17

Real and Complex Analysis

In the mid-1960s Walter Rudin (b. 1921), the author of a number of standard graduate textbooks in mathematics, wrote a textbook with the title *Real and Complex Analysis*, aimed at showing the considerable unity and overlap between the two subjects. It was necessary to write such a book because the two subjects, while sharing common roots in the calculus, had developed quite differently. The contrasts between the two are considerable. Complex analysis considers the smoothest, most orderly possible functions, those that are analytic, while real analysis allows the most chaotic imaginable functions. Complex analysis was, to pursue our botanical analogy, fully a “branch” of calculus, and foundational questions hardly entered into it. Real analysis had a share in both roots and branches, and it was intimately involved in the debate over the foundations of calculus.

What caused the two varieties of analysis to become so different? Both are dealing with functions, and both evolved under the stimulus of the differential equations of mathematical physics. The central point is the concept of a function. We have already seen the early definitions of this concept by Leibniz and Johann Bernoulli. All mathematicians from the seventeenth and eighteenth centuries had an intuitive picture of a function as a formula or expression in which variables are connected by rules derived from algebra or geometry. A function was regarded as continuous if it was given by a single formula throughout its range. If the formula changed, the function was called “mechanical” by Euler. Although “mechanical” functions may be continuous in the modern sense, they are not usually analytic. All the “continuous” functions in the older sense are analytic. They have power-series expansions, and those power-series expansions are often sufficient to solve differential equations. As a general signpost indicating where the paths diverge, the path of power-series expansions and the path of trigonometric-series expansions is a very good guide. A consequence of the development was that real-variable theory had to deal with very irregular and “badly behaved” functions. It was therefore in real analysis that the delicate foundational questions arose.

1. Complex analysis

Calculus began with a limited stock of geometry: a few curves and surfaces, all of which could be described analytically in terms of rational, trigonometric, exponential, and logarithmic functions of real variables. Soon, however, calculus was used to formulate problems in mathematical physics as differential equations. To solve those equations, the preferred technique was integration, but where integration failed, power series were the technique of first resort. These series automatically brought with them the potential of allowing the variables to assume complex values. But then integration and differentiation had to be suitably defined for complex functions of a complex variable. The result was a theory of analytic functions of a

complex variable whose range was much vaster than the materials that led to its creation.

In his 1748 *Introductio*, Euler emended the definition of a function, saying that a function is an *analytic expression* formed from a variable and constants. The rules for manipulating symbols were agreed on as long as only finite expressions were involved. But what did the symbols *represent*? Euler stated that variables were allowed to take on negative and imaginary values. Thus, even though the physical quantities the variables represented were measured as *positive rational* numbers, the algebraic and geometric properties of negative, irrational, and complex numbers could be invoked in the analysis. The extension from finite to infinite expressions was not long in coming. The extension of the calculus to complex numbers turned out to have monumental importance.

Lagrange undertook to reformulate the calculus in his treatises *Théorie des fonctions analytiques* (1797) and *Leçons sur le calcul des fonctions* (1801), basing it entirely on algebraic principles and stating as a fundamental premise that the functions to be considered are those that can be expanded in power series (having no negative or fractional powers of the variable). With this approach the derivatives of a function need not be defined as ratios of infinitesimals, since they can be defined in terms of the coefficients of the series that represents the function. Functions having a power series representation are known nowadays as *analytic functions* from the title of Lagrange's work.

1.1. Algebraic integrals. Early steps toward complexification were taken only on a basis of immediate necessity. As we have already seen, the applications of calculus in solving differential equations made the computation of integrals extremely important. Where computing the derivative never leads outside the class of elementary functions and leaves algebraic functions algebraic, trigonometric functions trigonometric, and exponential functions exponential, integrals are a very different matter. Algebraic functions often have nonalgebraic integrals, as Leibniz realized very early. The relation we now write as

$$\arccos(1-x) = \int_0^x \frac{1}{\sqrt{2t-t^2}} dt$$

was written by him as

$$a = \int dx : \sqrt{2x-x^2},$$

where $x = 1 - \cos a$.¹ Eighteenth-century mathematicians were greatly helped in handling integrals like this by the use of trigonometric functions. It was therefore natural that they would see the analogy when more complicated integrals came to be considered. Such problems arose from the study of pendulum motion and the rotation of solid bodies in physics, but we shall illustrate it with examples from pure geometry: the rectification of the ellipse and the division of the lemniscate into equal arcs. For the circle, we know that the corresponding problems lead to the integral

$$\int_0^x \frac{1}{\sqrt{1-t^2}} dt$$

¹ The limits of integration that we now use were introduced by Joseph Fourier in the nineteenth century.

for rectification and an equation

$$\int_0^y \frac{1}{\sqrt{1-t^2}} dt = \frac{1}{n} \int_0^x \frac{1}{\sqrt{1-t^2}} dt,$$

which can be written in differential form as

$$\frac{dx}{\sqrt{1-x^2}} = \frac{n dy}{\sqrt{1-y^2}},$$

for the division of an arc.

Trigonometry helps to solve this last equation. Instead of the arccosine function that the integral actually represents, it makes more sense to look at the inverse of it, the cosine function. This function provides an algebraic equation through its addition formula,

$$a_0 y^n - a_2 y^{n-2} + a_4 y^{n-4} - \cdots = x,$$

relating the abscissas of the end of the given arc (x) and the end of the n th part of it (y). The algebraic nature of this equation determines whether the division problem can be solved with ruler and compass. In particular, for $n = 3$ and a 60-degree arc ($x = 1/2$), for which the equation is $4y^3 - 3y = 1/2$, such a solution does not exist. Thus the problems of computing arc length on a circle and equal division of its arcs lead to an interesting combination of algebra, geometry, and calculus. Moreover, the periodicity of the inverse function makes this equation easy to solve (see Problem 17.1).

The division problem was fated to play an important role in study of integrals of algebraic functions. The Italian nobleman Fagnano (1682-1766) studied the problem of rectifying the lemniscate, whose polar equation is $r^2 = 2 \cos(2\theta)$. Its element of arc is $\sqrt{2}(1 - 2 \sin^2 \theta)^{-1/2} d\theta$, and the substitution $u = \tan \theta$ turns this expression into $\sqrt{2}(1 - u^4)^{-1/2} du$. Thus, the rectification problem involves evaluating the integral

$$\int_0^x \frac{\sqrt{2}}{\sqrt{1-u^4}} du,$$

while the division problem involves solving the differential equation

$$\frac{dz}{\sqrt{1-z^4}} = \frac{n du}{\sqrt{1-u^4}}.$$

Fagnano gave the solution for $n = 2$ as the algebraic relation

$$\frac{u\sqrt{2}}{\sqrt{1-u^4}} = \frac{1}{z} \sqrt{1 - \sqrt{1-z^4}}.$$

Euler observed the analogy between these integrals and the circular integrals just discussed, and suggested that it would be reasonable to study the inverse function. But Euler lived at a time when the familiar functions were still the elementary ones. He found a large number of integrals that could be expressed in terms of algebraic, logarithmic, and trigonometric functions and showed that there were others that could not be so expressed.

Legendre, Jacobi, and Abel. The foundation for further work in integration was laid by Legendre, who invented the term *elliptic integral*. Off and on for some 40 years between 1788 and 1828, he thought about integrals like those of Fagnano and Euler, classified them, computed their values, and studied their properties. He found their algebraic addition formulas and thereby reduced the division problem for these integrals to the solution of algebraic equations. Interestingly, he found that whereas the division problem requires solving an equation of degree n for the circle, it requires solving an equation of degree n^2 for the ellipse. After publishing his results as exercises in integral calculus in 1811, he wrote a comprehensive treatise in the 1820s. As he was finishing the third volume of this treatise he discovered a new set of transformations of elliptic integrals that made their computation easier. (He already knew one set of such transformations.) Just after the treatise appeared in 1827, he found to his astonishment that Jacobi had discovered the same transformations, along with others, and had connected them with the division problem. Jacobi's results in turn were partially duplicated by those of Abel.

Abel, who admired Gauss, was proud of having achieved the division of the lemniscate into 17 equal parts,² just as Gauss had done for the circle. The secret for the circle was to use the algebraic addition formula for trigonometric functions. For the lemniscate, as Legendre had shown, the equation was of higher degree. Abel was able to solve it by using complex variables, and in the process, he discovered that the inverse functions of the elliptic integrals, when regarded as functions of a complex variable, were *doubly periodic*. The double period accounted for the fact that the division equation was of degree n^2 rather than n . Without complex variables, the theory of elliptic integrals would have been a disconnected collection of particular results. With them, a great simplicity and unity was achieved. Abel went on to study algebraic addition formulas for very general integrals of the type

$$\int R(x, y(x)) dx,$$

where $R(x, y)$ is a rational function of x and y and $y(x)$ satisfies a polynomial equation $P(x, y(x)) = 0$. Such integrals are now called *Abelian integrals* in his honor. In particular, he established that for each polynomial $P(x, y)$ there was a number p , now called the *genus* of $P(x, y)$, such that a sum of any number of integrals $R(x, y)$ with different limits could be expressed in terms of just p integrals, whose limits of integration were rational functions of those in the given sum. For elliptic integrals, $p = 1$, and that is the content of the algebraic addition formulas discovered by Legendre. For a more complicated integral, say

$$\int \frac{1}{\sqrt{q(x)}} dx,$$

where $q(x)$ is a polynomial of degree 5 or higher, the genus may be higher. If $P(x, y) = y^2 - q(x)$, where the polynomial q is of degree $2p + 1$ or $2p + 2$, the genus is p .

After Abel's premature death, Jacobi continued to develop algebraic function theory. In 1832, he realized that for algebraic integrals of higher genus, the inverse functions could not be well defined, since there were p integrals and only one equation connecting them to the variable in terms of which they were to be expressed. He therefore had the idea of adjoining extra equations in order to get well-behaved

² Or, more generally, a Fermat prime number of parts.

inverses. For example, if $q(x)$ is of degree 5, he posed the problem of solving for x and y in terms of u and v in the equations

$$\begin{aligned} u &= \int_0^x \frac{1}{\sqrt{q(t)}} dt + \int_0^y \frac{1}{\sqrt{q(t)}} dt \\ v &= \int_0^x \frac{t}{\sqrt{q(t)}} dt + \int_0^y \frac{t}{\sqrt{q(t)}} dt. \end{aligned}$$

This problem became known as the *Jacobi inversion problem*. Solving it took a quarter of a century and led to progress in both complex analysis and algebra.

Jacobi himself gave it a start in connection with elliptic integrals. Although a nonconstant function that is analytic in the whole plane cannot be doubly periodic (because its absolute value cannot attain a maximum), a quotient of such functions can be, and Jacobi found the ideal numerators and denominators to use for expressing the doubly periodic elliptic functions as quotients: theta functions. The secret of solving the Jacobi inversion problem was to use theta functions in more than one complex variable, but working out the proper definition of those functions and the mechanics of applying them to the problem required the genius of Riemann and Weierstrass. These two giants of nineteenth-century mathematics solved the problem independently and simultaneously in 1856, but considerable preparatory work had been done in the meantime by other mathematicians. The importance of algebraic functions as the basic core of analytic function theory cannot be overemphasized. Klein (1926, p. 280) goes so far as to say that Weierstrass' purpose in life was

to conquer the inversion problem, even for hyperelliptic integrals of arbitrarily high order, as Jacobi had foresightedly posed it, perhaps even the problem for general Abelian integrals, using rigorous, methodical work with power series (including series in several variables).

It was in this way that the topic called the Weierstrass theory of analytic functions arose as a by-product.

1.2. Cauchy. Cauchy's name is associated most especially with one particular approach to the study of analytic functions of a complex variable, that based on complex integration. A complex variable is really two variables, as Cauchy was saying even as late as 1821. But a function is to be given by the same symbols, whether they denote real or complex numbers. When we integrate and differentiate a given function, which variable should we use? Cauchy discovered the answer, as early as 1814, when he first discussed such questions in print. The value of the function is also a pair of real numbers $u + iv$, and if the derivative is to be independent of the variable on which it is taken, these must satisfy the equations we now call the Cauchy-Riemann equations:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}; \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}.$$

In that case, as Cauchy saw, if we are integrating $u + iv$ in a purely formal way, separating real and imaginary parts, over a path from the lower left corner of a rectangle (x_0, y_0) to its upper right corner (x_1, y_1) , the same result is obtained whether the integration proceeds first vertically, then horizontally or first horizontally, then vertically. As Gauss had noted as early as 1811, Cauchy observed that the function

$1/(x + iy)$ did not have this property if the rectangle contained the point $(0, 0)$. The difference between the two paths was $2\pi i$, which Cauchy called the *residue*. Over the period from 1825 to 1840, Cauchy developed from this theorem what is now known as the Cauchy integral theorem, the Cauchy integral formula, Taylor's theorem, and the calculus of residues. The Cauchy integral theorem states that if γ is a curve enclosing a region in which $f(z)$ has a derivative then

$$\int_{\gamma} f(z) dz = 0.$$

If the real and imaginary parts of this integral are written out and compared with the Cauchy-Riemann equations, this formula becomes a simple consequence of what is known as Green's theorem (the two-variable version of the divergence theorem), published in 1828 by George Green (1793-1841) and simultaneously in Russia by Mikhail Vasilevich Ostrogradskii (1801-1862). When combined with the fact that the integral of $1/z$ around a curve that winds once around 0 is $2\pi i$, this theorem immediately yields as a consequence the Cauchy integral formula

$$f(z_0) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(z)}{z - z_0} dz.$$

When generalized, this formula becomes the residue theorem. Also from it, one can obtain estimates for the size of the derivatives. Finally, by expanding the denominator as a geometric series in powers of $z - z_1$, where z_1 lies inside the curve γ , one can obtain the Taylor series expansion of $f(z)$. These theorems form the essential core of modern first courses in complex analysis. This work was supplemented by a paper of Pierre Laurent (1813-1854), submitted to the Paris Academy in 1843, in which power series expansions about isolated singularities (Laurent series) were studied.

Cauchy was aware of the difficulties that arise in the case of multivalued functions and introduced the idea of a boundary curve (*ligne d'arrêt*) to prevent a function from assuming more than one value at a given point. As mentioned in Chapter 12, his student Puiseux studied the behavior of algebraic functions in the neighborhood of what we now call branch points, which are points c such that the function assumes many different values at each point of every neighborhood of c . Puiseux showed that at a branch point c near which there are n values of the function each of the n values of the function could be expanded in its own series of powers of a variable u such that $u^n = x - c$. The work of Cauchy, Laurent, Puiseux, and others thus brought complex analysis into existence as a well-articulated theory containing important principles and theorems.

1.3. Riemann. The work of Puiseux on algebraic functions of a complex variable was to be subsumed in two major papers of Riemann. The first of these, his doctoral dissertation, contained the concept now known as a Riemann surface. It was aimed especially at simplifying the study of an algebraic function $w(z)$ satisfying a polynomial equation $P(z, w(z)) \equiv 0$. In a sense, the Riemann surface revealed that all the significant information about the function was contained precisely in its singularities—the way it branched at its branch points. Information about the surface was contained in its *genus*, defined as half the total number of branch points,

counted according to order, less the number of sheets in the surface, plus 1.³ The Riemann surface of $w = \sqrt{z}$, for example, has two branch points (0 and ∞), each of order 1, and two sheets, resulting in genus 0. Riemann's geometric approach to the subject brought out the duality between surfaces and mappings of them, encapsulated in a formula known as the Riemann–Roch theorem (after Gustav Roch, 1839–1866). This formula connects the dimension of the space of functions on a Riemann surface having prescribed zeros and poles with the genus of the surface. Although it is a simple formula to write down, explaining the meaning of the terms in it requires considerable space, and so we omit the details.

In 1856 Riemann used his theory to give a very elegant solution of the Jacobi inversion problem. Since an analytic function must be constant if it has no poles on a Riemann surface, it was possible to use the periods of the integrals that occur in the problem to determine the function up to a constant multiple and then to find quotients of theta functions having the same periods, thereby solving the problem.

1.4. Weierstrass. Of the three founders of analytic function theory, Weierstrass was the most methodical. He had found his own solution to the Jacobi inversion problem and submitted it simultaneously with Riemann. When he saw Riemann's work, he withdrew his own paper and spent many years working out in detail how the two approaches related to each other. Where Riemann had allowed his geometric intuition to create castles in the air, so to speak, Weierstrass was determined to have a firm algebraic foundation. Instead of picturing kinematically a point wandering from one sheet of a Riemann surface to another, Weierstrass preferred a static object that he called a *Gebilde* (*structure*). His *Gebilde* was based on the set of pairs of complex numbers (z, w) satisfying a polynomial equation $p(z, w) = 0$, where $p(z, w)$ was an irreducible polynomial in the two variables. These pairs were supplemented by certain ideal points of the form (z, ∞) , (∞, w) , or (∞, ∞) when one or both of w or z tended to infinity as the other approached a finite or infinite value. Around all but a finite set of points, it was possible to expand w in an ordinary Taylor series in nonnegative integer powers of $z - z_0$. For each of the exceptional points, there would be one or more expansions in fractional or negative powers of $z - z_0$, as Puiseux and Laurent had found. These power series were Weierstrass' basic tool in analytic function theory.

Comparison of the three approaches. At first sight, it appears that Cauchy's approach, which is simultaneously analytic and geometric, subsumes the work of both Riemann and Weierstrass. Riemann, to be sure, had a more elegant way of overcoming the difficulty presented by multivalued functions, but Cauchy and Puiseux between them came very close to doing something logically equivalent. Weierstrass begins with the power series and considers only functions that have a power-series development, whereas Cauchy assumes only that the function is continuously differentiable.⁴ On the other hand, before you can verify Cauchy's basic assumption that a function is differentiable, you have to know what the function is. How is that information to be communicated, if not through some formula like a power series? Weierstrass saw this point clearly; in 1884 he said, "No matter how you twist and turn, you cannot avoid using some sort of analytic expressions" (quoted by Siegmund-Schultze, 1988, p. 253).

³ Klein (1926, p. 258) ascribes this definition to Alfred Clebsch (1833–1872).

⁴ It was shown by Edouard Goursat (1858–1936) in 1900 that differentiability implies continuous differentiability on open subsets of the plane.

2. Real analysis

In complex analysis attention is restricted from the outset to functions that have a complex derivative. That very strong assumption automatically ensures that the functions studied will have convergent Taylor series. If only mathematical physics could manage with just such smooth functions, the abstruse concepts that fill up courses in real analysis would not be needed. But the physical world is full of boundaries, where the density of matter is discontinuous, temperatures undergo abrupt changes, light rays reflect and refract, and vibrating membranes are clamped. For these situations the imaginary part of the variable, which often has no physical interpretation anyway, might as well be dropped, since its only mathematical role was to complete the analytic function. From that point on, analysis proceeds on the basis of real variables only. Real analysis, which represents another extension of calculus, has to deal with much more general and “rough” functions. All of the logical difficulties about calculus poured into this area of analysis, including the important questions of convergence of series, existence of maxima and minima, allowable ways of defining functions, continuity, and the meaning of integration. As a result, real analysis is so much less unified than complex analysis that it hardly appears to be a single subject. Its basic theorems do not follow from one another in any canonical order, and their proofs tend to be a bag of special tricks, rarely remembered for long except by professors who lecture on the subject constantly.

The free range of intuition suffered only minor checks in complex analysis. In that subject, what one wanted to believe very often turned out to be true. But real analysis almost seemed to be trapped in a hall of mirrors at times, as it struggled to gain the freedom to operate while avoiding paradoxes and contradictions. The generality of operations allowed in real analysis has fluctuated considerably over the centuries. While Descartes had imposed rather strict criteria for allowable curves (functions), Daniel Bernoulli attempted to represent very arbitrary functions as trigonometric series, and the mathematical physicist André-Marie Ampère (1775–1836) attempted to prove that a continuous function (in the modern sense, but influenced by preconceptions based on the earlier sense) would have a derivative at most points. The critique of this proof was followed by several decades of backtracking, as more and more exceptions were found for operations with series and integrals that appeared to be formally all right. Eventually, when a level of rigor was reached that eradicated the known paradoxes, the time came to reach for more generality. Georg Cantor’s set theory played a large role in this increasing generality, while developing paradoxes of its own. In the twentieth century, the theories of generalized functions and distributions restored some of the earlier freedom by inventing a new object to represent the derivative of functions that have no derivative in the ordinary sense.

2.1. Fourier series, functions, and integrals. There is a symmetry in the development of real and complex analysis. Broadly speaking, both arose from differential equations, and complex analysis grew out of power series, while real analysis grew out of trigonometric series. These two techniques, closely connected with each other through the relation $z^n = r^n(\cos n\theta + i \sin n\theta)$, led down divergent paths that nevertheless crossed frequently in their meanderings. The real and complex viewpoints in analysis began to diverge with the study of the vibrating string problem in the 1740s by d’Alembert, Euler, and Daniel Bernoulli.

For a string fastened at two points, say $(0, 0)$ and $(L, 0)$ and vibrating so that its displacement above or below the point $(x, 0)$ at time t is $y(x, t)$, mathematicians agreed that the best compromise between realism and comprehensibility to describe this motion was the *one-dimensional wave equation*, which d'Alembert studied in 1747,⁵ publishing the results in 1749:

$$\frac{\partial^2 y}{\partial t^2} = c^2 \frac{\partial^2 y}{\partial x^2}.$$

D'Alembert pointed out that the solution must be of the form

$$y(x, t) = \Psi(t + x) + \Gamma(t - x),$$

where for simplicity he assumed that $c = 1$. The equation alone does not determine the function, of course, since the vibrations depend on the initial position and velocity of the string. Accordingly, d'Alembert followed up with a prescribed initial position $f(x) = y(x, 0)$ and velocity $v(x) = \frac{\partial y}{\partial t}|_{t=0}$. He considered first the case when the initial position is identically zero, for which the function Ψ must be an even function of period $2L$, then the more general case.

The following year Euler took up this problem and commented on d'Alembert's solution. He observed that the initial position could be any shape at all, "either regular or irregular and mechanical." D'Alembert found that claim hard to accept. After all, the functions Ψ and Γ had to have periodicity and parity properties. How else could they be defined except as power series containing only odd or only even powers? Euler and d'Alembert were not interpreting the word "function" in the same way. Euler was even willing to consider initial positions $f(x)$ with corners (a "plucked" string), whereas d'Alembert insisted that $f(x)$ must have two derivatives simply to satisfy the equation.

Three years later Daniel Bernoulli tried to straighten this matter out, giving a solution in the form

$$y(x, t) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{L}\right) \cos\left(\frac{n\pi ct}{L}\right),$$

which he did not actually write out. Here the coefficients a_n were to be chosen so that the initial condition was satisfied, that is,

$$f(x) = y(x, 0) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{L}\right).$$

Observing that he had an infinite set of coefficients at his disposal for "fitting" the function, Bernoulli claimed that "any" function $f(x)$ had such a representation. Bernoulli's solution was the first of many instances in which the classical partial differential equations of mathematical physics—the wave, heat, and potential equations—were studied by separating variables and superposing the resulting solutions. The technique was ultimately to lead to what are called Sturm-Liouville problems, which we shall mention again below.

Before leaving the wave equation, we must mention one more important crossing between real and complex analysis in connection with it. In studying the action

⁵ Thirty years earlier Brook Taylor (1685–1731) had analyzed the problem geometrically and concluded that the normal acceleration at each point would be proportional to the normal curvature at that point. That statement is effectively the same as this equation, and was quoted by d'Alembert.

of gravity, Pierre-Simon Laplace (1749–1827) was led to what is now known as Laplace's equation in three variables. The two-variable version of this equation is

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0.$$

The operator on the left-hand side of this equation is known as the *Laplacian*. Since Laplace's equation can be thought of as the wave equation with velocity $c = \sqrt{-1}$, complex numbers again enter into a physical problem. Recalling d'Alembert's solution of the wave equation, Laplace suggested that the solutions of his equation might be sought in the form $f(x + y\sqrt{-1}) + g(x - y\sqrt{-1})$. Once again a problem that started out as a real-variable problem led inexorably to the need to study functions of a complex variable.

The definition of a function. Daniel Bernoulli accepted his father's definition of a function as "an expression formed in some manner from variables and constants," as did Euler and d'Alembert. But those words seemed to have different meanings for each of them. Daniel Bernoulli thought that his solution met the criterion of being "an expression formed from variables and constants." His former colleague in the Russian Academy of Sciences,⁶ Euler, saw the matter differently. This time it was Euler who argued that the concept of function was being used too loosely. According to him, since the right-hand side of Bernoulli's formula consisted of odd functions of period $2L$, it could represent only an odd function of period $2L$. Therefore, he said, it did not have the generality of the solution he and d'Alembert had given. Bottazzini (1986, p. 29) expresses the situation very well, saying, "We are here facing a misunderstanding that reveals one aspect of the contradictions between the old and new theory of functions, even though they are both present in the same man, Euler, the protagonist of this transformation." The difference between the old and new concepts is seen in the simplest example, the function $|x|$, which equals x when $x \geq 0$ and $-x$ for $x \leq 0$. We have no difficulty thinking of this function as one function. It appeared otherwise to nineteenth-century mathematicians. Fourier described what he called a "discontinuous function represented by a definite integral" in 1822: the function

$$\frac{2}{\pi} \int_0^\infty \frac{\cos qx}{1+q^2} dq = \begin{cases} e^{-x} & \text{if } x \geq 0, \\ e^x & \text{if } x \leq 0. \end{cases}$$

Fifty years later Gaston Darboux (1844–1918) gave the modern point of view, that this function is not truly discontinuous but merely a function expressed by two different analytic expressions in different parts of its domain.

The change in point of view came about gradually, but an important step was Cauchy's refinement of the definition in the first chapter of his 1821 *Cours d'analyse*:

When variable quantities are related so that, given the value of one of them, one can infer those of the others, we normally consider that the quantities are all expressed in terms of one of them, which is called the *independent* variable, while the others are called *dependent variables*.

⁶ Bernoulli had left St. Petersburg in 1733, Euler in 1741.

Cauchy's definition still does not specify what *ways* of expressing one variable in terms of another are legitimate, but this definition was a step toward the basic idea that the value of the independent variable determines (uniquely) the value of the dependent variable or variables.

Fourier series. Daniel Bernoulli's work introduced trigonometric series as an alternative to power series. In his classic work of 1811, a revised version of which was published in 1821,⁷ *Théorie analytique de chaleur* (*Analytic Theory of Heat*), Fourier established the standard formulas for the Fourier coefficients of a function. For an even function of period 2π , these formulas are

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos nx; \quad a_n = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos nx \, dx, \quad n = 0, 1, \dots$$

A trigonometric series whose coefficients are obtained from an integrable function $f(x)$ in this way is called a *Fourier series*.

After trigonometric series had become a familiar technique, mathematicians were encouraged to look for other simple functions in terms of which solutions of more general differential equations than Laplace's equation could be expressed. Between 1836 and 1838 this problem was attacked by Charles Sturm (1803–1855) and Joseph Liouville, who considered general second-order differential equations of the form

$$[p(x)y'(x)]' + [\lambda r(x) + q(x)]y(x) = 0.$$

When a solution of Laplace's equation is sought in the form of a product of functions of one variable, the result is often an equation of this type for the one-variable functions. It often happens that only isolated values of λ yield solutions satisfying given boundary conditions. Sturm and Liouville found that in general there will be an infinite set of values $\lambda = \lambda_n$, $n = 1, 2, \dots$, satisfying the equation and a pair of conditions at the endpoints of an interval $[a, b]$, and that these values increase to infinity. The values can be arranged so that the corresponding solutions $y_n(x)$ have exactly n zeros in $[a, b]$, and any solution of the differential equation can be expressed as a series

$$y(x) = \sum_{n=1}^{\infty} c_n y_n(x).$$

The sense in which such series converge was still not clear, but it continued to be studied by other mathematicians. It required some decades for all these ideas to be sorted out clearly.

Proving that a Fourier series actually did converge to the function that generated it was one of the first places where real analysis encountered greater difficulties than complex analysis. In 1829 Dirichlet proved that the Fourier series of $f(x)$ converged to $f(x)$ for a bounded periodic function $f(x)$ having only a finite number of discontinuities and a finite number of maxima and minima in each period.⁸ Dirichlet tried to get necessary and sufficient conditions for convergence, but that is a problem that has never been solved. He showed that some kind of continuity would be required by giving the famous example of the function whose value at x is one of two different values according as x is rational or irrational. This function

⁷ The original version remained unpublished until 1972, when Grattan-Guinness published an annotated version of it.

⁸ We would call such a function *piecewise monotonic*.

is called the *Dirichlet function*. For such a function, he thought, no integral could be defined, and therefore no Fourier series could be defined.⁹

Fourier integrals. The convergence of the Fourier series of $f(x)$ can be expressed as the equation

$$f(x) = \frac{1}{\pi} \int_0^\pi f(y) dy + \frac{2}{\pi} \sum_{n=1}^{\infty} \int_0^\pi f(y) \cos(ny) \cos(nx) dy.$$

That equation may have led to an analogous formula for Fourier integrals, which appeared during the early nineteenth century in papers on the wave and heat equations written by Poisson, Laplace, Fourier, and Cauchy. The central discovery in this area was the Fourier inversion formula, which we now write as

$$f(x) = \frac{2}{\pi} \int_0^\infty \int_0^\infty f(y) \cos(zy) \cos(zx) dy dz.$$

The analogy with the formula for series is clear: The continuous variable z replaces the discrete index n , and the integral on z replaces the sum over n . Once again, the validity of the representation is much more questionable than the validity of the formulas of complex analysis, such as the Cauchy integral formula for an analytic function. The Fourier inversion formula has to be interpreted very carefully, since the order of integration cannot be reversed. If the integrals make sense in the order indicated, that happy outcome can only be the result of some special properties of the function $f(x)$. But what are those properties?

The difficulty was that the integral extended over an infinite interval so that convergence required the function to have two properties: It needed to be continuous, and it needed to decrease sufficiently rapidly at infinity to make the integral converge. These properties turned out to be, in a sense, dual to each other. Considering just the inner integral as a function of z :

$$\hat{f}(z) = \int_0^\infty f(y) \cos(zy) dy,$$

it turns out that the more rapidly $f(y)$ decreases at infinity, the more derivatives $\hat{f}(z)$ has, and the more derivatives $f(y)$ has, the more rapidly $\hat{f}(z)$ decreases at infinity. The converses are also, broadly speaking, true. Could one insist on having both conditions, so that the representation would be valid? Would these assumptions impair the usefulness of these techniques in mathematical physics? Alfred Pringsheim (1850-1941, father-in-law of the great writer Thomas Mann) studied the Fourier integral formula (1910), noting especially the two kinds of conditions that $f(x)$ needed to satisfy, which he called “conditions in the finite region” (“im Endlichen”) and “conditions at infinity” (“im Unendlichen”). Nowadays, they are called local and global conditions. Pringsheim noted that the local conditions could be traced all the way back to Dirichlet’s work of 1829, but said that “a rather obvious backwardness reveals itself” in regard to the global conditions.

⁹ The increasing latitude allowed in analysis, mentioned above, is illustrated very well by this example. When the Lebesgue integral is used, this function is regarded as identical with the constant value it assumes on the irrational numbers.

[They] seem in general to be limited to a relatively narrow condition, one which is insufficient for even the simplest type of application, namely that of absolute integrability of $f(x)$ over an infinite interval. There are, as far as I know, only a few exceptions.

Thus, to the question as to whether physics could get by with sufficiently smooth functions $f(x)$ that decay sufficiently rapidly, the answer turned out to be, in general, no. Physics needs to deal with discontinuous integrable functions $f(y)$, and for these $f(z)$ cannot decay rapidly enough at infinity to make its integral converge, at least not absolutely. What was to be done?

One solution involved the introduction of *convergence factors*, leading to a more general sense of convergence, called Abel–Poisson convergence. In a paper on wave motion published in 1818 Poisson used the representation

$$f(x) = \frac{1}{\pi} \int_0^\infty \int_{-\infty}^{+\infty} f(\alpha) \cos \alpha(x - \alpha) e^{-k\alpha} d\alpha da.$$

The exponential factor provided enough decrease at infinity to make the integral converge. Poisson claimed that the resulting integral tended toward $f(x)$ as k decreased to 0.

Abel used a similar technique to justify the natural value assigned to nonabsolutely convergent series such as

$$\ln(2) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots \quad \text{and} \quad \frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots.$$

which can be obtained by expanding the integrands of the following integrals as geometric series and integrating termwise:

$$\int_0^1 \frac{1}{1+r} dr; \quad \int_0^1 \frac{1}{1+r^2} dr.$$

In Abel's case, the motive for making a careful study of continuity was his having noticed that a trigonometric series could represent a discontinuous function. From Paris in 1826 he wrote to a friend that the expansion

$$\frac{x}{2} = \sin x - \frac{1}{2} \sin 2x + \frac{1}{3} \sin 3x - \frac{1}{4} \sin 4x + \cdots$$

was provable for $0 \leq x < \pi$, although obviously it could not hold at $x = \pi$. Thus, while the representation might be a good thing, it meant, on the other hand, that the sum of a series of continuous functions could be discontinuous. Abel also believed that many of the difficulties mathematicians were encountering were traceable to the use of divergent series. He gave, accordingly, a thorough discussion of the convergence of the binomial series, the most difficult of the elementary Taylor series to analyze.¹⁰

For the two conditionally convergent series shown above and the general Fourier integral, continuity of the sum was needed. In both cases, what appeared to be a necessary evil—the introduction of the convergence factor $e^{-k\alpha}$ or r —turned out to have positive value. For the functions $r^n \cos n\theta$ and $r^n \sin n\theta$ are harmonic functions

¹⁰ Unknown to Abel, Bolzano had discussed the binomial series in 1816, considering integer, rational, and irrational (real) exponents, admitting that he could not cover all possible cases, due to the incomplete state of the theory of complex numbers at the time (Bottazzini, 1986, pp. 96–97). He performed a further analysis of series in general in 1817, with a view to proving the intermediate value property (see Section 4 of Chapter 12).

if r and θ are regarded as polar coordinates, while $e^{-ay} \cos(ax)$ and $e^{-ay} \sin(ax)$ are harmonic if x and y are regarded as rectangular coordinates. The factor used to ensure convergence was providing harmonic functions, at no extra cost.

General trigonometric series. The study of trigonometric functions advanced real analysis once again in 1854, when Riemann was required to give a lecture to qualify for the position of *Privatdocent* (roughly what would be an assistant professor nowadays). As the rules required, he was to propose three topics and the faculty would choose the one he lectured on. One of the three, based on conversations he had had with Dirichlet over the preceding year, was the representation of functions by trigonometric series.¹¹ Dirichlet was no doubt hoping for more progress toward necessary and sufficient conditions for convergence of a Fourier series, the topic he had begun so promisingly a quarter-century earlier. Riemann concentrated on one question in particular: *Can a function be represented by more than one trigonometric series?* That is, *can two trigonometric series with different coefficients have the same sum at every point?* In the course of his study, Riemann was driven to examine the fundamental concept of integration. Cauchy had defined the integral

$$\int_a^b f(x) dx$$

as the number approximated by the sums

$$\sum_{n=1}^N f(x_n)(x_n - x_{n-1})$$

as N becomes large, where $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$. Riemann refined the definition slightly, allowing $f(x_n)$ to be replaced by $f(x_n^*)$ for any x_n^* between x_{n-1} and x_n . The resulting integral is known as the Riemann integral today. Riemann sought necessary and sufficient conditions for such an integral to exist. The condition that he formulated led ultimately to the concept of a set of measure zero,¹² half a century later: *For each $\sigma > 0$ the total length of the intervals on which the function $f(x)$ oscillates by more than σ must become arbitrarily small if the partition is sufficiently fine.*

2.2. Completeness of the real numbers. The concept now known as completeness of the real numbers is associated with the *Cauchy convergence criterion*, which asserts that a sequence of real numbers $\{a_n\}_{n=1}^\infty$ converges to some real number a if it is a *Cauchy sequence*; that is, for every $\varepsilon > 0$ there is an index n such that $|a_n - a_k| < \varepsilon$ for all $k \geq n$. This condition was stated somewhat loosely by Cauchy in his *Cours d'analyse*, published in the mid-1820s, and the proof given there was also somewhat loose. The same criterion had been stated, and for sequences of functions rather than sequences of numbers, a decade earlier by Bolzano.

¹¹ As the reader will recall from Chapter 12, this topic was *not* the one Riemann did lecture on. Gauss preferred the topic of foundations of geometry, and so Riemann's paper on trigonometric series was not published until 1867, after his death.

¹² A set of points on the line has measure zero if for every $\varepsilon > 0$ it can be covered by a sequence of intervals (a_k, b_k) whose total length is less than ε .

2.3. Uniform convergence and continuity. Cauchy was not aware at first of any need to make the distinction between pointwise and uniform convergence, and he even claimed that the sum of a series of continuous functions would be continuous, a claim contradicted by Abel, as we have seen. The distinction is a subtle one. It is all too easy not to notice whether choosing n large enough to get a good approximation when $f_n(x)$ converges to $f(x)$ requires one to take account of which x is under consideration. That point was rather difficult to state precisely. The first clear statement of it is due to Philipp Ludwig von Seidel (1821–1896), a professor at Munich, who in 1847 studied the examples of Dirichlet and Abel, coming to the following conclusion:

When one begins from the certainty thus obtained that the proposition cannot be generally valid, then its proof must basically lie in some still hidden supposition. When this is subject to a precise analysis, then it is not difficult to discover the hidden hypothesis. One can then reason backwards that this [hypothesis] cannot occur [be fulfilled] with series that represent discontinuous functions. [Quoted in Bottazzini, 1986, p. 202]

In order to reason confidently about continuity, derivatives, and integrals, mathematicians began restricting themselves to cases where the series converged uniformly. Weierstrass, in particular, provided a famous theorem known as the M -test for uniform convergence of a series. But, although the M -test is certainly valuable in dealing with power series, uniform convergence in general is too severe a restriction. The important trigonometric series studied by Abel, for example, represented a discontinuous function as the sum of a series of continuous functions and therefore did not converge uniformly. Yet it could be integrated term by term. One could provide many examples of series of continuous functions that converged to a continuous function but not uniformly. Weaker conditions were needed that would justify the operations rigorously without restricting their applicability too strongly.

2.4. General integrals and discontinuous functions. The search for less restrictive hypotheses and the consideration of more general figures on a line than just points and intervals led to more general notions of length, area, and integral, allowing more general functions to be integrated. Analysts began generalizing the integral beyond the refinements introduced by Riemann. Foundational problems also added urgency to this search. For example, in 1881, Vito Volterra (1860–1940) gave an example of a continuous function having a derivative at every point, but whose derivative was not Riemann integrable. What could the fundamental theorem of calculus mean for such a function, which had an antiderivative but no integral, as integrals were then understood?

New integrals were created by the Latvian mathematician Axel Harnack (1851–1888), by the French mathematicians Émile Borel (1871–1956), Henri Lebesgue (1875–1941), and Arnaud Denjoy (1884–1974), and by the German mathematician Oskar Perron (1880–1975). By far the most influential of these was the Lebesgue integral, which was developed between 1899 and 1902. This integral was to have profound influence in the area of probability, due to its use by Borel, and in trigonometric series representations, an application that Lebesgue developed, perhaps as proof of the usefulness of his highly abstruse integral, which, as a former colleague of the author was fond of saying, “did not change any tables of integrals.” Lebesgue

justified his more general integral in the following words, from the preface to his 1904 monograph.

[I]f we wished to limit ourselves always to these good [that is, smooth] functions, we would have to give up on the solution of a number of easily stated problems that have been open for a long time. It was the solution of these problems, rather than a love of complications, that caused me to introduce in this book a definition of the integral that is more general than that of Riemann and contains the latter as a special case.

Despite its complexity—to develop it with proofs takes four or five times as long as developing the Riemann integral—the Lebesgue integral was included in textbooks as early as 1907: for example, *Theory of Functions of a Real Variable*, by E. W. Hobson (1856–1933). Its chief attraction was the greater generality of the conditions under which it allowed termwise integration. Following the typical pattern of development in real analysis, the Lebesgue integral soon generated new questions. The Hungarian mathematician Frigyes Riesz (1880–1956) introduced the classes now known as L_p -spaces, the spaces of measurable functions¹³ f for which $|f|^p$ is Lebesgue integrable, $p > 0$. (The space L_∞ consists of functions that are bounded on a set whose complement has measure zero.) How the Fourier series and integrals of functions in these spaces behave became a matter of great interest, and a number of questions were raised. For example, in his 1915 dissertation at the University of Moscow, Nikolai Nikolaevich Luzin (1883–1950) posed the conjecture that the Fourier series of a square-integrable function converges except on a set of measure zero. Fifty years elapsed before this conjecture was proved by the Swedish mathematician Lennart Carleson (b. 1928).

2.5. The abstract and the concrete. The increasing generality allowed by the notation $y = f(x)$ threatened to carry mathematics off into stratospheric heights of abstraction. Although Ampère had tried to show that a continuous function is differentiable at most points, the attempt was doomed to failure. Bolzano constructed a “sawtooth” function in 1817 that was continuous, yet had no derivative at any point. Weierstrass later used an absolutely convergent trigonometric series to achieve the same result,¹⁴ and a young Italian mathematician Salvatore Pincherle (1853–1936), who took Weierstrass’ course in 1877–1888, wrote a treatise in 1880 in which he gave a very simple example of such a function (Bottazzini, 1986, p. 286):

$$f(x) = \sum_{n=1}^{\infty} \frac{\sin(n!x)}{n!}.$$

Volterra’s example of a continuous function whose derivative was not integrable, together with the examples of continuous functions having no derivative at any point naturally cast some doubt on the applicability of the abstract concept of continuity and even the abstract concept of a function. Besides the construction of more general integrals and the consequent ability to “measure” more complicated

¹³ See below for the definition.

¹⁴ This example was communicated by his student Paul du Bois-Reymond (1831–1889) in 1875. The following year du Bois-Reymond constructed a continuous periodic function whose Fourier series failed to converge at a set of points that came arbitrarily close to every point.

geometric figures, it was necessary to investigate differentiation in more detail as well.

The secret of that quest turned out to be not continuity, but monotonicity. A continuous function may fail to have a derivative, but in order to fail, it must oscillate very wildly, as the examples of Bolzano and Weierstrass did. A function that did not oscillate or oscillated only a finite total amount, necessarily had a derivative except on a set of measure zero. The ultimate result in this direction was achieved by Lebesgue, who showed that a monotonic function has a derivative on a set whose complement has measure zero. Such a function might or might not be the integral of its derivative, as the fundamental theorem of calculus states. In 1902 Lebesgue gave necessary and sufficient conditions for the fundamental theorem of calculus to hold; a function that satisfies these conditions, and is consequently the integral of its derivative, is called *absolutely continuous*.

To return to the problem of abstractness, we note that it had been known at least since the time of Lagrange that any finite set of n data points (x_k, y_k) , $k = 1, \dots, n$, with x_k all different, could be fitted perfectly with a polynomial of degree at most $n - 1$. Such a polynomial might—indeed, probably would—oscillate wildly in the intervals between the data points. Weierstrass showed in 1884 that any continuous function, no matter how abstract, could be uniformly approximated by a polynomial over any bounded interval $[a, b]$. Since there is always some observational error in any set of data, this result meant that polynomials could be used in both practical and theoretical ways, to fit data, and to establish general theorems about continuous functions. Weierstrass also proved a second version of the theorem, for periodic functions, in which he showed that for these functions the polynomial could be replaced by a finite sum of sines and cosines. This connection to the classical functions freed mathematicians to use the new abstract functions, confident that in applications they could be replaced by computable functions.

Weierstrass lived before the invention of the new abstract integrals mentioned above arose, although he did encourage the development of the abstract set theory of Georg Cantor on which these integrals were based. With the development of the Lebesgue integral a new category of functions arose, the *measurable functions*. These are functions $f(x)$ such that the set of x for which $f(x) > c$ always has a meaningful measure, although it need not be a geometrically simple set, as it is in the case of continuous functions. It appeared that Weierstrass' work needed to be repeated, since his approximation theorem did not apply to measurable functions. In his 1915 dissertation Luzin produced two beautiful theorems in this direction. The first was what is commonly called by his name nowadays, the theorem that for every measurable function $f(x)$ and every $\varepsilon > 0$ there is a continuous function $g(x)$ such that $g(x) \neq f(x)$ only on a set of measure less than ε . As a consequence of this result and Weierstrass' approximation theorem, it followed that every measurable function is the limit of a sequence of polynomials on a set whose complement has measure zero. Luzin's second theorem was that every finite-valued measurable function is the *derivative* of a continuous function at the points of a set whose complement has measure zero. He was able to use this result to show that any prescribed set of measurable boundary values on the disk could be the boundary values of a harmonic function.

With the Weierstrass approximation theorem and theorems like those of Luzin, modern analysis found some anchor in the concrete analysis of the "classical" period that ran from 1700 to 1900. But that striving for generality and freedom of

operation still led to the invocation of some strong principles of inference in the context of set theory. By mid-twentieth century mathematicians were accustomed to proving concrete facts using abstract techniques. To take just one example, it can be proved that some differential equations have a solution because a contraction mapping of a complete metric space must have a fixed point. Classical mathematicians would have found this proof difficult to accept, and many twentieth-century mathematicians have preferred to write in “constructivist” ways that avoid invoking the abstract “existence” of a mathematical object that cannot be displayed explicitly. But most mathematicians are now comfortable with such reasoning.

2.6. Discontinuity as a positive property. The Weierstrass approximation theorems imply that the property of being the limit of a sequence of continuous functions is no more general than the property of being the limit of a sequence of polynomials or the sum of a trigonometric series. That fact raises an obvious question: What kind of function is the limit of a sequence of continuous functions? As noted above, du Bois-Reymond had shown that it can be discontinuous on a set that is, as we now say, dense. But can it, for example, be discontinuous at *every* point? That was one of the questions that interested René-Louis Baire (1874–1932). If one thinks of discontinuity as simply the absence of continuity, classifying mathematical functions as continuous or discontinuous seems to make no more sense than classifying mammals as cats or noncats. Baire, however, looked at the matter differently. In his 1905 *Leçons sur les fonctions discontinues* (*Lectures on Discontinuous Functions*) he wrote

Is it not the duty of the mathematician to begin by studying in the abstract the relations between these two concepts of continuity and discontinuity, which, while mutually opposite, are intimately connected?

Strange as this view may seem at first, we may come to have some sympathy for it if we think of the dichotomy between the continuous and the discrete, that is, between geometry and arithmetic. At any rate, to a large number of mathematicians at the turn of the twentieth century, it did not seem strange. The Moscow mathematician Nikolai Vasilevich Bugaev (1837–1903, father of the writer Andrei Belyi) was a philosophically inclined scholar who thought it possible to establish two parallel theories, one for continuous functions, the other for discontinuous functions. He called the latter theory *arithmology* to emphasize its arithmetic character. There is at least enough of a superficial parallel between integrals and infinite series and between continuous and discrete probability distributions (another area in which Russia has produced some of the world’s leaders) to make such a program plausible. It is partly Bugaev’s influence that caused works on set theory to be translated into Russian during the first decade of the twentieth century and brought the Moscow mathematicians Luzin and Dmitrii Fyodorovich Egorov (1869–1931) and their students to prominence in the area of measure theory, integration, and real analysis.

Baire’s monograph was single-mindedly dedicated to the pursuit of one goal: to give a necessary and sufficient condition for a function to be the pointwise limit of a sequence of continuous functions. He found the condition, building on earlier ideas introduced by Hermann Hankel (1839–1873): The necessary and sufficient condition is that the discontinuities of the function form a set of *first category*.

A set is of first category if it is the union of a sequence of sets A_k such that every interval (a, b) contains an interval (c, d) disjoint from A_k . All other sets are of second category.¹⁵ Although interest in the specific problems that inspired Baire has waned, the importance of his work has not. The whole edifice of what is now functional analysis rests on three main theorems, two of which are direct consequences of what is called the Baire category theorem (that a complete metric space is of second category as a subset of itself) and cannot be proved without it. Here we have an example of an unintended and fortuitous consequence of one bit of research turning out to be useful in an area not considered by its originator.

Questions and problems

17.1. The familiar formula $\cos \theta = 4 \cos^3(\theta/3) - 3 \cos(\theta/3)$, can be rewritten as $p(\cos \theta/3, \cos \theta) = 0$, where $p(x, y) = 4x^3 - 3x - y$. Observe that $\cos(\theta + 2m\pi) = \cos \theta$ for all integers m , so that

$$p\left(\cos\left(\frac{\theta + 2m\pi}{3}\right), \cos \theta\right) \equiv 0,$$

for all integers m . That makes it very easy to construct the roots of the equation $p(x, \cos \theta) = 0$. They must be $\cos((\theta + 2m\pi)/3)$ for $m = 0, 1, 2$. What is the analogous equation for dividing a circular arc into five equal pieces?

Suppose (as is the case for elliptic integrals) that the inverse function of an integral is doubly periodic, so that $f(x + m\omega_1 + n\omega_2) = f(x)$ for all m and n . Suppose also that there is a polynomial $p(x)$ of degree n^2 such that $p(f(\theta/n)) = f(\theta)$. Show that the roots of the equation $p(x) = f(\theta)$ must be $f(\theta/n + (k/n)\omega_1 + (l/n)\omega_2)$, where k and l range independently from 0 to $n - 1$.

17.2. Show that if $y(x, t) = (f(x + ct) + f(x - ct))/2$ is a solution of the one-dimensional wave equation that is valid for all x and t , and $y(0, t) = 0 = y(L, t)$ for all t , then $f(x)$ must be an odd function of period $2L$.

17.3. Show that the problem $X''(x) - \lambda X(x) = 0$, $Y''(y) + \lambda Y(y) = 0$, with boundary conditions $Y(0) = Y(2\pi)$, $Y'(0) = Y'(2\pi)$, implies that $\lambda = n^2$, where n is an integer, and that the function $X(x)Y(y)$ must be of the form $(c_n e^{nx} + d_n e^{-nx})(a_n \cos(ny) + b_n \sin(ny))$ if $n \neq 0$.

17.4. Show that the differential equation

$$\frac{dx}{\sqrt{1-x^4}} + \frac{dy}{\sqrt{1-y^4}} = 0$$

has the solution $y = [(1 - x^2)/(1 + x^2)]^{1/2}$. Find another obvious solution of this equation.

17.5. Show that Fourier series can be obtained as the solutions to a Sturm-Liouville problem on $[0, 2\pi]$ with $p(x) = r(x) \equiv 1$, $q(x) = 0$, with the boundary conditions $y(0) = y(2\pi)$, $y'(0) = y'(2\pi)$. What are the possible values of λ ?

¹⁵ In his work on set theory, discussed in Section 4 of Chapter 12, Hausdorff criticized this terminology as "colorless."

The History of Mathematics: A Brief Course, Second Edition

by Roger Cooke

Copyright © 2005 John Wiley & Sons, Inc.

Part 7

Mathematical Inferences

At various points in this survey of mathematics we have found mathematicians debating the meaning of what they were doing and the legitimacy of their procedures. In this last part of our study we examine the ways in which mathematics enters into the process of *drawing conclusions*. Human beings draw conclusions with differing levels of confidence, based on experience. At the one extreme are opinions about the most complicated phenomena around us, other human beings and human society as a whole. These matters are so complex that very few statements about them can be simultaneously free of significant doubt and of great importance. At the other end are matters that are so simple and obvious that we do not hesitate to label as insane anyone who doubts them. At the very least, we do not bother refuting a person who says that the word *yes* requires 10 letters to spell or that the city of Tuscaloosa is located in Siberia. Mathematics itself lies on the “confident” end of the spectrum, but its applications to practical life do not always share that certainty. Chapter 18 surveys the uncertain, in the form of the history of probability and statistics. Chapter 19 covers the more certain, in the form of the history of logic.

CHAPTER 18

Probability and Statistics

The need to make decisions on the basis of incomplete data is very widespread in human life. We need to decide how warmly to dress and whether to carry an umbrella when we leave home in the morning. We may have to decide whether to risk a dangerous but potentially life-saving medical procedure. Such decisions rely on statistical reasoning. Statistics is a science that is not exactly mathematics. It *uses* mathematics, in the form of probability, but its procedures are the inverse ones of fitting probability distributions to real-world data. Probability theory, on the other hand, is a form of pure mathematics, with theorems that are just as certain as those in algebra and analysis. We begin this chapter with the pure mathematics and end with its application.

1. Probability

The word *probability* is related to the English words *probe*, *probation*, *prove*, and *approve*. All of these words originally had a sense of *testing* or *experimenting*,¹ reflecting their descent from the Latin *probo*, which has these meanings. In other languages the word used in this mathematical sense has a meaning more like *plausibility*,² as in the German *Wahrscheinlichkeit* (literally, *truth resemblance*) or the Russian *veroyatnost'* (literally, *credibility*, from the root *ver-*, meaning *faith*). The concept is very difficult to define in declarative sentences, precisely because it refers to phenomena that are normally described in the subjunctive mood. This mood has nearly disappeared in modern English; it clings to a precarious existence in the past tense, "If it were true that..." having replaced the older "If it be true that...". The language of Aristotle and Plato, however, who were among the first people to discuss chance philosophically, had two such moods, the subjunctive and the optative, sometimes used interchangeably. As a result, they could express more easily than we the intuitive concepts involved in discussing events that are imagined rather than observed.

Intuitively, probability attempts to express the relative strength of the feeling of confidence we have that an event will occur. How surprised would we be if the event happened? How surprised would we be if it did not happen? Because we do have different degrees of confidence in certain future events, quantitative concepts become applicable to the study of probability. Generally speaking, if an event occurs essentially all the time under specified conditions, such as an eclipse

¹ The common phrase "the exception that proves the rule" is nowadays misunderstood and misused because of this shift in the meaning of the word *prove*. Exceptions *test* rules, they do not *prove* them in the current sense of that word. In fact, quite to the contrary, exceptions *disprove* rules.

² Here is another interesting word etymology. The root is *plaudo*, meaning *strike*, but specifically meaning to clap one's hands together, to applaud. Once again, *approval* is involved in the notion of probability.

of the Sun, we use a deterministic model (geometric astronomy, in this case) to study and predict it. If it occurs sometimes under conditions frequently associated with it, we rely on probabilistic models. Some earlier scientists and philosophers regarded probability as a measure of our ignorance. Kepler, for example, believed that the supernova of 1604 in the constellation Serpent may have been caused by a random collision of particles; but in general he was a determinist who thought that our uncertainty about a roll of dice was merely a matter of insufficient data being available. He admitted, however, that he could find no law to explain the apparently random pattern of eccentricities in the elliptical orbits of the six planets known to him.

Once the mathematical subject got started, however, it developed a life of its own, in which theorems could be proved with the same rigor as in any other part of mathematics. Only the application of those theorems to the physical world remained and remains clouded by doubt. We use probability informally every day, as the weather forecast informs us that the chance of rain is 30% or 80% or 100%,³ or when we are told that one person in 30 will be afflicted with Alzheimer's disease between the ages of 65 and 74. Much of the public use of such probabilistic notions is, although not meaningless, at least irrelevant. For example, we are told that the life expectancy of an average American is now 77 years. Leaving aside the many questionable assumptions of environmental and political stability used in the model that produced this fascinating number, we should at least ask one question: Can the number be related to the life of any person in any meaningful way? What plans can one base on it, since anyone may die on any given day, yet very few people can confidently rule out the possibility of living past age 90?⁴

The many uncertainties of everyday life, such as the weather and our health, occur mixed with so many possibly relevant variables that it would be difficult to distill a theory of probability from those intensely practical matters. What is needed is a simpler and more abstract model from which principles can be extracted and gradually made more sophisticated. The most obvious and accessible such models are games of chance. On them probability can be given a quantitative and empirical formulation, based on the frequency of wins and losses. At the same time, the imagination can arrange the possible outcomes symmetrically and in many cases assign equal probabilities to different events. Finally, since money generally changes hands at the outcome of a game, the notion of a random variable (payoff to a given player, in this case) as a quantity assuming different values with different probabilities can be modeled.

1.1. Cardano. The systematic mathematization of probability began in sixteenth-century Italy with Cardano. Cardano gambled frequently with dice and attempted to count the favorable cases for a throw of three dice. His table of values, as reported by Todhunter (1865, p. 3) is as follows.

³ These numbers are generated by computer models of weather patterns for squares in a grid representing a geographical area. The modeling of their accuracy also uses probabilistic notions (see Problem 18.1).

⁴ The Russian mathematician Yu. V. Chaikovskii (2001) believes that some of this cloudiness is about to be removed with the creation of a new science he calls *aleatics* (from the Latin word *alea*, meaning *dice-play* or *gambling*). We must wait and see. A century ago, other Russian mathematicians confidently predicted a bright future for "arithmology." Prophecy is the riskiest of all games of chance.

1	2	3	4	5	6	7	8	9	10	11	12
108	111	115	120	126	133	33	36	37	36	33	26

Readers who enjoy playing with numbers may find some amusement here. Since it is impossible to roll a 1 with three dice, the table value should perhaps be interpreted as the number of ways in which 1 may appear on *at least* one of the three dice. If so, then Cardan has got it wrong. One can imagine him thinking that if a 1 appears on one of the dice, the other two may show 36 different numbers, and since there are three dice on which the 1 may appear, the total number of ways of rolling a 1 must be $3 \cdot 36$ or 108. That way of counting ignores the fact that in some of these cases 1 appears on two of the dice or all three. By what is now known as the inclusion-exclusion principle, the total should be $3 \cdot 36 - 3 \cdot 6 + 1 = 91$. But it is difficult to say what Cardano had in mind. The number 111 given for 2 may be the result of the same count, increased by the three ways of choosing two of the dice to show a 1. Todhunter worked out a simple formula giving these numbers, but could not imagine any gaming rules that would correspond to them. If indeed Cardano made mistakes in his computations, he was not the only great mathematician to do so.

Cardano's *Liber de ludo* (*Book on Gambling*) was published about a century after his death. In this book Cardano introduces the idea of assigning a probability p between 0 and 1 to an event whose outcome is not certain. The principal applications of this notion were in games of chance, where one might bet, for example, that a player could roll a 6 with one die given three chances. The subject is not developed in detail in Cardano's book, much of which is occupied by descriptions of the actual games played. However, Cardano does state the multiplicative rule for a run of successes in independent trials. Thus the probability of getting a six on each of three successive rolls with one die is $(\frac{1}{6})^3$. Most important, he recognized the real-world application of what we call the law of large numbers, saying that when the probability for an event is p , then after a large number n of repetitions, the number of times it will occur does not lie far from the value np . This law says that it is not certain that the number of occurrences will be near np , but "that is where the smart money bets."

After a bet has been made and before it is settled, a player cannot unilaterally withdraw from the bet and recover her or his stake. On the other hand, an accountant computing the net worth of one of the players ought to count part of the stake as an asset owned by that player; and perhaps the player would like the right to sell out and leave the game. What would be a fair price to charge someone for taking over the player's position? More generally, what happens if the game is interrupted? How are the stakes to be divided? The principle that seemed fair was that, *regardless of the relative amount of the stake each player had bet, at each moment in the game a player should be considered as owning the portion of the stakes equal to that player's probability of winning at that moment*. Thus, the net worth of each player is constantly changing as the game progresses, in accordance with what we now call *conditional probability*. Computing these probabilities in games of chance usually involves the combinatorial counting techniques the reader has no doubt encountered.

1.2. Fermat and Pascal. A French nobleman, the Chevalier de Méré, who was fond of gambling, proposed to Pascal the problem of dividing the stakes in a game where one player has bet that a six will appear in eight rolls of a single die, but

the game is terminated after three unsuccessful tries. Pascal wrote to Fermat that the player should be allowed to sell the throws one at a time. If the first throw is foregone, the player should take one-sixth of the stake, leaving five-sixths. Then if the second throw is also foregone, the player should take one-sixth of the remaining five-sixths or $\frac{5}{36}$, and so on. In this way, Pascal argued that the fourth through eighth throws were worth $\frac{1}{6}[(\frac{5}{6})^3 + (\frac{5}{6})^4 + (\frac{5}{6})^5 + (\frac{5}{6})^6 + (\frac{5}{6})^7]$.

This expression is the value of those throws *before* any throws have been made. If, after the bets are made but before any throws of the die have been made, the bet is changed and the players agree that only three throws shall be made, then the player holding the die should take this amount as compensation for sacrificing the last five throws. Remember, however, that the net worth of a player is constantly changing as the game progresses and the probability of winning changes. The value of the fourth throw, for example, is smaller to begin with, since there is some chance that the player will win before it arrives, in which case it will not arrive. At the beginning of the game, the chance of winning on the fourth roll is $(\frac{5}{6})^3 \frac{1}{6}$, the factor $(\frac{5}{6})^3$ representing the probability that the player will *not* have won before then. After three unsuccessful throws, however, the probability that the player "will not have" won (because he *did not* win) on the first three throws is 1, and so the probability of winning on the fourth throw becomes $\frac{1}{6}$.

Fermat expressed the matter as follows:

[T]he three first throws having gained nothing for the player who holds the die, the total sum thus remaining at stake, he who holds the die and who agrees not to play his fourth throw should take $\frac{1}{6}$ as his reward. And if he has played four throws without finding the desired point and if they agree that he shall not play the fifth time, he will, nevertheless, have $\frac{1}{6}$ of the total for his share. Since the whole sum stays in play it not only follows from the theory, but it is indeed common sense that each throw should be of equal value.

Pascal wrote back to Fermat, proclaiming himself satisfied with Fermat's analysis and overjoyed to find that "the truth is the same at Toulouse and at Paris."

1.3. Huygens. Huygens wrote a treatise on probability in 1657. His *De ratiociniis in ludo aleæ* (*On Reasoning in a Dice Game*) consisted of 14 propositions and contained some of the results of Fermat and Pascal. In addition, Huygens was able to consider multinomial problems, involving three or more players. Cardano's idea of an *estimate of the expectation* was elaborated by Huygens. He asserted, for example, that if there are p (equally likely) ways for a player to gain a and q ways to gain b , then the player's expectation is $(pa + qb)/(p + q)$.

Even simple problems involving these notions can be subtle. For example, Huygens considered two players A and B taking turns rolling the dice, with A going first. Any time A rolls a 6, A wins; any time B rolls a 7, B wins. What are the relative chances of winning? (The answer to that question would determine the fair proportions of the stakes to be borne by the two players.) Huygens concluded that the odds were 31:30 in favor of B , that is, A 's probability of winning was $\frac{30}{61}$ and B 's probability was $\frac{31}{61}$.

1.4. Leibniz. Although Leibniz wrote a full treatise on combinatorics, which provides the mathematical apparatus for computing many probabilities in games of chance, he did not himself gamble. But he did analyze many games of chance and suggest modifications of them that would make them fair (zero-sum) games. Some of his manuscripts on this topic have been analyzed by de Mora-Charles (1992). One of the games he analyzed is known as quinquenove. This game is played between two players using a pair of dice. One of the players, called the banker, rolls the dice, winning if the result is either a double or a total number of spots showing equal to 3 or 11. There are thus 10 equally likely ways for the banker to win with this roll, out of 36 equally likely outcomes. If the banker rolls a 5 or 9 (hence the name “quinquenove”), the other player wins. The other player has eight ways of winning of the equally likely 36 outcomes, leaving 18 ways for the game to end in a draw. The reader will be fascinated and perhaps relieved to learn that the great Leibniz, author of *De arte combinatoria*, confused permutations and combinations in his calculations for this game and got the probabilities wrong.

1.5. The *Ars Conjectandi* of Jakob Bernoulli. One of the classic founding documents of probability theory was published in 1713, eight years after the death of its author, Leibniz’ disciple Jakob Bernoulli. This work, *Ars conjectandi* (*The Art of Prediction*), moved probability theory beyond the limitations of analyzing games of chance. It was intended by its author to apply mathematical methods to the uncertainties of life. As he said in a letter to Leibniz, “I have now finished the major part of the book, but it still lacks the particular examples, the principles of the art of prediction that I teach how to apply to society, morals, and economics. . . .” That was an ambitious undertaking, and Bernoulli had not quite finished the work when he died in 1705.

Bernoulli gave a very stark picture of the gap between theory and application, saying that only in simple games such as dice could one apply the equal-likelihood approach of Fermat and Pascal, whereas in the cases of interest, such as human health and longevity, no one had the power to construct a suitable model. He recommended statistical studies as the remedy to our ignorance, saying that if 200 people out of 300 of a given age and constitution were known to have died within 10 years, it was a 2-to-1 bet that any other person of that age and constitution would die within a decade.

In this treatise Bernoulli reproduced the problems solved by Huygens and gave his own solution of them. He considered what are now called *Bernoulli trials* in his honor. These are repeated experiments in which a particular outcome either happens (success) with probability b/a or does not happen (failure) with probability c/a , the same probability each time the experiment is performed, each outcome being independent of all others. (A simple nontrivial example is rolling a single die, counting success as rolling a 5. Then the probabilities are $\frac{1}{6}$ and $\frac{5}{6}$.) Since $b/a + c/a = 1$, Bernoulli saw correctly that the binomial expansion, and hence Pascal’s triangle, would be useful in computing the probability of getting at least m successes in n trials. He gave that probability as

$$\sum_{k=m}^n \binom{n}{k} \left(\frac{b}{a}\right)^k \left(\frac{c}{a}\right)^{n-k}.$$

It was, incidentally, in this treatise, when computing the sum of the c th powers of the first n integers, that Bernoulli introduced what are now called the *Bernoulli*

numbers, defined by the formula

$$\sum_{k=1}^n k^c = \frac{n^{c+1}}{c+1} + \frac{n}{2} + \frac{c}{2}An^{c-1} + \frac{c(c-1)(c-2)}{2 \cdot 3 \cdot 4}Bn^{c-3} + \dots$$

Nowadays we define these numbers as $B_0 = 1$, $B_1 = -\frac{1}{2}$, and thence $B_2 = A$, $B_3 = -B$, and so forth. He illustrated his formula by finding

$$\sum_{k=1}^{1000} k^{10} = 91409924241424243424241924242500.$$

The law of large numbers. Bernoulli imagined an urn containing numbers of black and white pebbles, whose ratio is to be determined by sampling with replacement. Here it is possible that you will always get a white pebble, no matter how many times you sample. However, if black pebbles constitute a significant proportion of the contents of the urn, this outcome is very unlikely. After discussing the degree of certainty that would suffice for practical purposes (he called it *virtual certainty*),⁵ he noted that this degree of certainty could be attained empirically by taking a sufficiently large sample. The probability that the empirically determined ratio would be close to the true ratio increases as the sample size increases, but the result would be accurate only within certain limits of error. More precisely, given certain limits of tolerance, by a sufficient number of trials,

[W]e can attain any desired degree of probability that the ratio found by our many repeated observations will lie between these limits.

This last assertion is an informal statement of the law of large numbers for what are now called *Bernoulli trials*, that is, repeated independent trials with the same probability of a given outcome at each trial. If the probability of the outcome is p and the number of trials is n , this law can be phrased precisely by saying that for any $\varepsilon > 0$ there exists a number n_0 such that if m is the number of times the outcome occurs in n trials and $n > n_0$, the probability that the inequality $|(m/n) - p| > \varepsilon$ will hold is less than ε .⁶ Bernoulli stated this principle in terms of the segment of the binomial series of $(r + s)^{n(r+s)}$ consisting of the n terms on each side of the largest term (the term containing $r^n s^n$), and he proved it by giving an estimate on n sufficient to make the ratio of this sum to the sum of the remaining terms at least c , where c is specified in advance. This problem is the earliest in which probability and statistics were combined to solve a problem of practical application.

⁵ This phrase is often translated more literally as *moral certainty*, which has the wrong connotation.

⁶ Probabilists say that the frequency of successes converges "in probability" to the probability of success at each trial. Analysts say it converges "in measure." There is also a strong law of large numbers, more easily stated in terms of independent random variables, which asserts that (under suitable hypotheses) there is a set of probability 1 on which the convergence to the mean occurs. That is, the convergence is "almost surely," as probabilists say and "almost everywhere," as analysts phrase the matter. On a finite measure space such as a probability space, almost everywhere convergence implies convergence in measure, but the converse is not true.

1.6. De Moivre. In 1711, even before the appearance of Jakob Bernoulli's treatise, another groundbreaking book on probability appeared, the *Doctrine of Chances*, written by Abraham de Moivre (1667–1754), a French Huguenot who took refuge in England after 1685, when Louis XIV revoked the Edict of Nantes, which had guaranteed civil rights for Huguenots when Henri IV took the French throne in 1598.⁷ De Moivre's book went through several editions. Its second edition, which appeared in 1738, introduced a significant piece of numerical analysis, useful for approximating sums of terms of a binomial expansion $(a + b)^n$ for large n . De Moivre had published the work earlier in a paper written in 1733. Having no notation for the base e , which was introduced by Euler a few years later, de Moivre simply referred to the hyperbolic (natural) logarithm and "the number whose logarithm is 1." De Moivre first considered only the middle term of the expansion. That is, for an even power $n = 2m$, he estimated the term

$$\binom{2m}{m} = \frac{(2m)!}{(m!)^2}$$

and found it equal to $\frac{2}{B\sqrt{n}}$, where B was a constant for which he knew only an infinite series. At that point, he got stuck, as he admitted, until his friend James Stirling (1692–1770) showed him that "the Quantity B did denote the Square-root of the Circumference of a Circle whose Radius is Unity." In our terms, $B = \sqrt{2\pi}$, but de Moivre simply wrote c for B . Without having to know the exact value of B de Moivre was able to show that "the Logarithm of the Ratio, which a Term distant from the middle by the Interval l , has the the middle Term, is [approximately, for large n] $-\frac{2l^2}{n}$." In modern language,

$$\binom{2n}{n+l} / \binom{2n}{n} \approx e^{-2l^2/n}.$$

De Moivre went on to say, "The Number, which answers to the Hyperbolic Logarithm $-2ll/n$, [is]

$$1 - \frac{2l}{n} + \frac{4l^2}{2nn} - \frac{8l^3}{6n^3} + \frac{16l^4}{24n^4} - \frac{32l^5}{120n^5} + \frac{64l^6}{720n^6}, \text{ \&c.} "$$

By scaling, de Moivre was able to estimate segments of the binomial distribution. In particular, the fact that the numerator was l^2 and the denominator n allowed him to estimate the probability that the number of successes in Bernoulli trials would be between fixed limits. He came close to noticing that the natural unit of probability for n trials was a multiple of \sqrt{n} . In 1893 this natural unit of measure for probability was named the *standard deviation* by the British mathematician Karl Pearson (1857–1936). For Bernoulli trials with probability of success p at each trial the standard deviation is $\sigma = \sqrt{np(1-p)}$.

For what we would call a coin-tossing experiment in which $p = \frac{1}{2}$ —he imagined tossing a metal disk painted white on one side and black on the other—de Moivre observed that with 3600 coin tosses, the odds would be more than 2 to 1 against a deviation of more than 30 "heads" from the expected number of 1800. The standard deviation for this experiment is exactly 30, and 68 percent of the area under a normal curve lies within one standard deviation of the mean. De Moivre

⁷ The spirit of sectarianism has infected historians to the extent that Catholic and Protestant biographers of de Moivre do not agree on how long he was imprisoned in France for being a Protestant. They do agree that he was imprisoned, however. To be fair to the French, they did elect him a member of the Academy of Sciences a few months before his death.

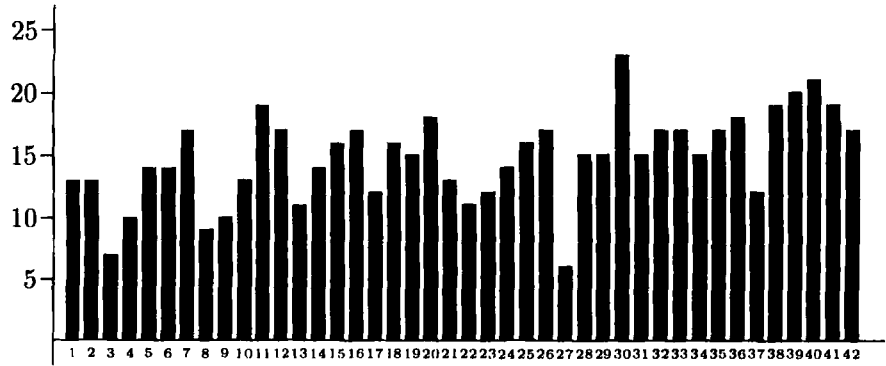


FIGURE 1. Frequencies of numbers in a state lottery over a one-year period.

could imagine the bell-shaped normal curve that we are familiar with, but he could not give it an equation. Instead he described it as the curve whose ordinates were numbers having certain logarithms. What seems most advanced in his analysis is that he recognized the area under the curve as a probability and computed it by a mechanical quadrature method that he credited jointly to Newton, Roger Cotes, James Stirling, and himself. This tendency of the average of many independent trials to look like the bell-shaped curve is called the *central limit theorem*.

It is difficult to appreciate the work of Bernoulli and de Moivre in applications without seeing it applied in a real-world illustration. To take a very simple example, consider Fig. 1, which is a histogram of the frequencies with which the numbers from 1 to 42 were drawn in a state lottery over a period of one year⁸. Six numbers are drawn twice a week, for a total of 624 numbers each year. At each drawing a given number has a probability of $\frac{1}{7}$ of being drawn. Thus, focusing attention only on the occurrence of a fixed integer k , we can think of the lottery as a series of 104 independent trials with a probability of success (drawing the number k) equal to $\frac{1}{7}$ at each trial.

Although the individual data do not reveal the binomial distribution or show any bell-shaped curve, we can think of the frequencies with which the 42 numbers are drawn as the data for a second probabilistic model. By the binomial distribution, for each frequency r from 0 to 104, The probability that a given number will be drawn r times should theoretically be

$$\binom{104}{r} \left(\frac{1}{7}\right)^r \left(\frac{6}{7}\right)^{104-r}.$$

If the probability of an event is proportional to the number of times that the event occurs in a large number of trials, then the number of numbers drawn r times should be 42 times this expression. The resulting theoretical frequencies are negligibly small for $r < 6$ or $r > 23$. The values predicted by this theoretical model for r between 6 and 23, rounded to the nearest integer, are given in the second row of the following table, while the experimentally observed numbers are given in the bottom row.

⁸ The Tri-state Megabucks of Maine, New Hampshire, and Vermont, from mid-December 2000 to mid-December 2001.

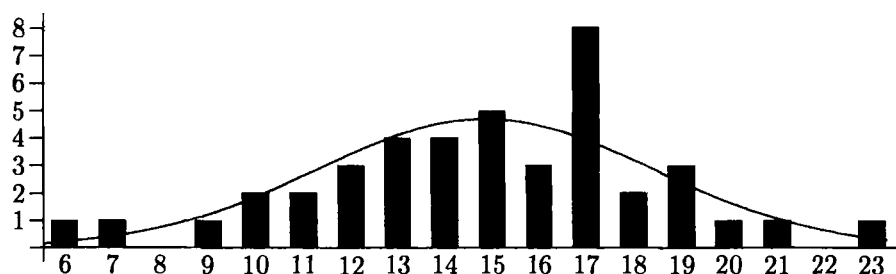


FIGURE 2. Histogram of the frequencies of the frequencies in Fig. 1, compared with a normal distribution.

Freq.	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Pred.	0	0	1	1	2	3	4	4	5	5	4	4	3	2	2	1	1	0
Obs.	1	1	0	1	2	2	3	4	4	5	3	8	2	3	1	1	0	1

The agreement is not perfect, nor would we expect it to be. But it is remarkably close, except for the one “outlier” at a frequency of 17, attained by eight numbers instead of the theoretically predicted four. The mean for this model is $104/7 \approx 14.85$, and the standard deviation is $\sqrt{624/7} \approx 3.569$. The histogram for the frequencies of the frequencies, compared with the graph of the standard bell-shaped curve with this mean and standard deviation are shown in Fig. 2. The fact that mere numerical reasoning *compels* even the most chaotic phenomena to exhibit some kind of order is one of the most awe-inspiring aspects of applied probability theory. It is the phenomenon that led the British mathematician Francis Galton (1822–1911) to describe the normal distribution as “the supreme law of unreason.”

The Petersburg paradox. Soon after its introduction by Huygens and Jakob Bernoulli the concept of mathematical expectation came in for some critical appraisal. While working in the Russian Academy of Sciences, Daniel Bernoulli discussed the problem now known as the *Petersburg paradox* with his brother Niklaus (1695–1726, known as Niklaus II). We can describe this paradox informally as follows. Suppose that you toss a coin until heads appears. If it appears on the first toss, you win \$2, if it first appears on the second toss, you win \$4, and so on; if heads first appears on the n th toss, you win 2^n dollars. How much money would you be willing to pay to play this game? Now by “rational” computations the expected winning is infinite, being $2 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} + 8 \cdot \frac{1}{8} + \dots$, so that you should be willing to pay, say, \$10,000 to play each time. On the other hand, who would bet \$10,000 knowing that there was an even chance of winning back only \$2, and that the odds are 7 to 1 against winning more than \$10? Something more than mere expectation was involved here. Daniel Bernoulli discussed the matter at length in an article in the *Comentarii* of the Petersburg Academy for 1730–1731 (published in 1738). He argued for the importance of something that we now call *utility*. If you already possess an amount of money x and you receive a small additional amount of money dx , how much *utility* does the additional money have for you, subjectively? Bernoulli assumed that the increment of utility dy was directly proportional to dx and inversely proportional to x , so that

$$dy = \frac{k dx}{x},$$

and as a result, the total utility of personal wealth is a logarithmic function of total wealth. One consequence of this assumption is a law of diminishing returns: The additional satisfaction from additional wealth decreases as wealth increases. Bernoulli used this idea to explain why any rational person would refuse to play the game. Obviously, the expected gain in utility from each of these wins, being proportional to the logarithm of the money gained, has a finite total, and so one should be willing to pay only an amount of money that has an equal utility to the gambler. A different explanation can be found in Problem 18.4 below. This explanation seems to have been given first by the mathematician John Venn (1834–1923) of Caius⁹ College, Cambridge in 1866.

The utility y , which Bernoulli called the *emolumentum* (*gain*), is an important tool in economic analysis, since it provides a dynamic model of economic behavior: Buyers exchange money for goods or services of higher personal utility; sellers exchange goods and services for money of higher personal utility. If money, goods, and services did not have different utility for different people, no market could exist at all.¹⁰ That idea is valid independently of the actual formula for utility given by Bernoulli, although, as far as measurements of psychological phenomena can be made, Bernoulli's assumption was extremely good. The physiologist Ernst Heinrich Weber (1795–1878) asked blindfolded subjects to hold weights, which he gradually increased, and to say when they noticed an increase in the weight. He found that the threshold for a noticeable difference was indeed inversely proportional to the weight. That is, if S is the perceived weight and W the actual weight, then $dS = k dW/W$, where dW is the smallest increment that can be noticed and dS the corresponding perceived increment. Thus he found exactly the law *assumed* by Bernoulli for perceived increases in wealth.¹¹ Utility is of vital importance to the insurance industry, which makes its profit by having a large enough stake to play “games” that resemble the Petersburg paradox.

Mathematically, there was an important concept missing from the explanation of the Petersburg paradox. Granted that one should expect the “expected” value of a quantity depending on chance, how *confidently* should one expect it? The question of *dispersion* or *variance* of a random quantity lies beneath the surface here and needed to be brought out. It turns out that when the expected value is infinite, or even when the variance is infinite, no rational projections can be made. However, since we live in a world of finite duration and finite resources, each “game” will be played only a finite number of times. It follows that every actual game has a finite expectation and variance and is subject to rational analysis using them.

1.7. Laplace. Although Laplace is known primarily as an astronomer, he developed a great deal of theoretical physics. (The differential equation satisfied by harmonic functions is named after him.) He also understood the importance of probabilistic methods for processing the results of measurements. In his *Théorie analytique des probabilités*, he proved that the distribution of the average of random observational errors that are uniformly distributed in an interval symmetric about zero tends to the normal distribution as the number of observations increases.

⁹ Pronounced “Keys.”

¹⁰ One feels the lack of this concept very strongly in the writing on economics by Aristotle and his followers, especially in their condemnation of the practice of lending money at interest.

¹¹ Weber's result was publicized by Gustave Theodor Fechner (1801–1887) and is now known as the Weber–Fechner law.

Except for using the letter c where we now use e to denote the base of natural logarithms, he had what we now call the central limit theorem for independent uniformly distributed random variables.

1.8. Legendre. In a treatise on ways of determining the orbits of comets, published in 1805, Legendre dealt with the problem that frequently results when observation meets theory. Theory prescribes a certain number of equations of a certain form to be satisfied by the observed quantities. These equations involve certain theoretical parameters that are not observed, but are to be determined by fitting observations to the theoretical model. Observation provides a large number of empirical, approximate solutions to these equations, and thus normally provides a number of equations far in excess of the number of parameters to be chosen. If the law is supposed to be represented by a straight line, for example, only two constants are to be chosen. But the observed data will normally not lie on the line; instead, they may cluster around a line. How is the observer to choose canonical values for the parameters from the observed values of each of the quantities?

Legendre's solution to this problem is now a familiar technique. If the theoretical equation is $y = f(x)$, where $f(x)$ involves parameters α, β, \dots , and one has data points (x_k, y_k) , $k = 1, \dots, n$, sum the squares of the "errors" $f(x_k) - y_k$ to get an expression in the parameters

$$E(\alpha, \beta, \dots) = \sum_{k=1}^n (f(x_k) - y_k)^2,$$

and then choose the parameters so as to minimize E . For fitting with a straight line $y = ax + b$, for example, one needs to choose $E(a, b)$ given by

$$E(a, b) = \sum_{k=1}^n (ax_k + b - y_k)^2$$

so that

$$\frac{\partial E}{\partial a} = 0 = \frac{\partial E}{\partial b}.$$

1.9. Gauss. Legendre was not the first to tackle the problem of determining the most likely value of a quantity x using the results of repeated measurements of it, say x_k , $k = 1, \dots, n$. In 1799 Laplace had tried the technique of taking the value x that minimizes the sum of the absolute errors¹² $|x - x_k|$. But still earlier, in 1794 as shown by his diary and correspondence, the teenager Gauss had hit on the least-squares technique for the same purpose. However, as Reich (1977, p. 56) points out, Gauss did not consider this discovery very important and did not publish it until 1809. In 1816 Gauss published a paper on observational errors, in which he discussed the most probable value of a variable based on a number of observations of it. His discussion was much more modern in its notation than those that had gone before, and also much more rigorous. He found the likelihood of an error of size x to be

$$\frac{h}{\sqrt{\pi}} e^{-h^2 x^2},$$

where h was what he called the *measure of precision*. He showed how to estimate this parameter by inverse-probability methods. In modern terms, $1/\sqrt{2h}$ is the

¹² This method has the disadvantage that one large error and many small errors count equally. The least-squares technique avoids that problem.

standard deviation. This work brought the normal distribution into a more or less standard form, and it is now often referred to as the *Gaussian distribution*.

1.10. Philosophical issues. The notions of chance and necessity have always played a large role in philosophical speculation; in fact, most books on logic are kept in the philosophy sections of libraries. Many of the mathematicians who have worked in this area have had a strong interest in philosophy and have speculated on what probability means. In so doing, they have come up against the same difficulties that confront natural philosophers when trying to explain how induction works. Granted that like Pavlov's dogs and Skinner's pigeons (see Chapter 1), human beings tend to form expectations based on frequent, but not necessarily invariable conjunctions of events and seem to find it very difficult to suspend judgment and live with no belief where there is no evidence,¹³ can philosophy offer us any assurance that proceeding by induction based on probability and statistics is any better than, say, divination such as one finds in the *I Ching*? Are insurance companies acting on *pure faith* when they offer to bet us that we will survive long enough to pay them more money in premiums than they will pay out when we die? If probability is a subjective matter, is subjectivity the same as arbitrariness?

What, then, is probability, when applied to the physical world? Is it merely a matter of frequency of observation, and consequently objective? Or do human beings have some innate faculty for assigning probabilities? For example, when we toss a coin twice, there are four distinguishable outcomes: HH, HT, TH, TT. Are these four equally likely? If one does not know the order of the tosses, only three possibilities can be distinguished: two heads, two tails, and one of each. Should those be regarded as equally likely, or should we imagine that we do know the order and distinguish all four possibilities?¹⁴ Philosophers still argue over such matters. Siméon-Denis Poisson (1781–1840) seemed to be having it both ways in his *Recherches sur la probabilité des jugemens (Investigations into the Plausibility of Inferences)* when he wrote that

The probability of an event is the reason we have to believe that it has taken place, or that it will take place.

and then immediately followed up with

The measure of the probability of an event is the ratio of the number of cases favorable to that event, to the total number of cases favorable or contrary.

In the first statement, he appeared to be defining probability as a subjective event, one's own *personal* reason, but then proceeded to make that reason an objective thing by assuming equal likelihood of all outcomes. Without some restriction on the universe of discourse, these definitions are not very useful. We do not know, for example, whether our automobile will start tomorrow morning or not, but if

¹³ In his *Formal Logic*, Augustus de Morgan imagined asking a person selected at random for an opinion whether the volcanoes—he meant craters—on the unseen side of the moon were larger than those on the side we can see. He concluded, "The odds are, that though he has never thought of the question, he has a pretty stiff opinion in three seconds."

¹⁴ If the answer to that question seems intuitively obvious, please note that in more exotic applications of statistics, such as in quantum mechanics, either possibility can occur. Fermions have wave functions that are antisymmetric, and they distinguish between HT and TH; bosons have symmetric wave functions and do not distinguish them.

the probability of its doing so were really only 50% because there are precisely two possible outcomes, most of us would not bother to buy an automobile. Surely Poisson was assuming some kind of symmetry that would allow the imagination to assign equal likelihoods to the outcomes, and intending the theory to be applied only in those cases. Still, in the presence of ignorance of causes, equal probabilities seem to be a reasonable starting point. The law of entropy in thermodynamics, for example, can be deduced as a tendency for an isolated system to evolve to a state of maximum probability, and maximum probability means the maximum number of equally likely states for each particle.

1.11. Large numbers and limit theorems. The idea of the law of large numbers was stated imprecisely by Cardano and with more precision by Jakob Bernoulli. To better carry out the computations involved in using it, de Moivre was led to approximate the binomial distribution with what we now realize was the normal distribution. He, Laplace, and Gauss all grasped with different degrees of clarity the principle (central limit theorem) that when independent measurements are averaged, they tend to shape themselves into the bell-shaped curve.

The law of large numbers was given its name in the 1837 work of Poisson just mentioned. Poisson discovered an approximation to the probability of getting at most k successes in n trials, valid when n is large and the probability p is small. He thereby introduced what is now known as the *Poisson distribution*, in which the probability of k successes is given by

$$p_k = e^{-\lambda} \frac{\lambda^k}{k!}.$$

The Russian mathematician Chebyshev introduced the concept of a random variable and its mathematical expectation. He is best known for his 1846 proof of the weak law of large numbers for repeated independent trials. That is, he showed that the probability that the actual proportion of successes will differ from the expected proportion by less than any specified $\varepsilon > 0$ tends to 1 as the number of trials increases. In 1867 he proved what is now called *Chebyshev's inequality*: *The probability that a random variable will assume a value more than [what is now called] k standard deviations from its mean is at most $1/k^2$.* This inequality was published by Chebyshev's friend and translator Irénée-Jules Bienaymé (1796–1878) and is sometimes called the *Chebyshev–Bienaymé inequality* (see Heyde and Seneta, 1977). This inequality implies the weak law of large numbers. In 1887 Chebyshev also gave an explicit statement of the central limit theorem for independent random variables.

The extension of the law of large numbers to dependent trials was achieved by Chebyshev's student Andrei Andreevich Markov (1856–1922). The subject of dependent trials—known as *Markov chains*—remains an object of current research. In its simplest form it applies to a system in one of a number of states $\{S_1, \dots, S_n\}$ which at specified times may change from one state to another. If the probability of a transition from S_i to S_j is p_{ij} , the matrix

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{pmatrix}$$

is called the *transition matrix*. If successive transitions are all independent of one another, one can easily verify that the matrix power P^k gives the probabilities of the transitions in k steps.

2. Statistics

The subject of probability formed the theoretical background for the empirical science known as statistics. Some theoretical analysis of the application of probability to hypothesis testing and modification is due to Thomas Bayes (1702–1761), a British clergyman. Bayes' articles were published in 1764–1765 (after his death) by Rev. Richard Price (1723–1791). Bayes considered the problem opposite to that considered by Jakob Bernoulli. Where Bernoulli assigned probabilities to the event of getting k successes in n independent trials, assuming the probability of success in each trial was p , Bayes analyzed the problem of finding the probability p based on an observation that k successes and $n - k$ failures have occurred. In other words, he tried to estimate the parameter in a distribution from observed data. His claim was that p would lie between a and b with a probability proportional to the area under the curve $y = x^k(1 - x)^{n-k}$ between those limits. He then analyzed a more elaborate example. Suppose we know that the probability of event B is p , given that event A has occurred. Suppose also that, after a number of trials, without reference to whether A has occurred or not, we find that event B has occurred m times and has not occurred n times. What probability should be assigned to A ? Bayes' example of event A was a line drawn across a billiard table parallel to one of its sides at an unknown distance x from the left-hand edge. A billiard ball is rolled at random across the table, coming to rest on the left of the line m times and on the right of it n times. Assuming that the width of the table is a , the probability of the ball resting left of the line is x/a , and the probability that it rests on the right is $1 - x/a$. How can we determine x from the actual observed frequencies m and n ? Bayes' answer was that the probability that x lies between b and c is proportional to the area under the curve $y = x^m(a - x)^n$ between those two values. This first example of statistical estimation is also the first *maximum-likelihood* estimation, since the "density" function $x^m(a - x)^n$ has its maximum value where $m(a - x) = nx$, that is, $x = a \frac{m}{m+n}$, so that the proportion $m : n = x : a - x$ holds. It seems intuitively reasonable that the most likely value of x is the value that makes this proportion correct, and that the likelihood decreases as x moves away from this value. This is the kind of reasoning used by Gauss in his 1816 paper on the estimation of observational errors to find the parameter (measure of precision) in the normal distribution. To derive this result, Bayes had to introduce the concept of conditional probability. The probability of A , *given that* B has occurred, is equal to the probability that both events happen divided by the probability of B . (If B has occurred, it must have positive probability, and therefore the division is legitimate.) Although Bayes stated this much with reasonable clarity (see Todhunter, 1865, p. 298), the full statement of what is now called Bayes' theorem (see below) is difficult to discern in his analysis.

The word *statistics* comes from the state records of births, deaths, and other economic facts that governments have always found it necessary to keep for administrative purposes. The raw data form far too large a set of numbers to be analyzed individually in most cases, and that is where probabilistic models and inverse-probability reasoning, such as that used by Bayes and Gauss become most

useful. An early example was an argument intended to prove that the world was designed for human habitation based on the ratio of male to female births. In 1710 Queen Anne's physician John Arbuthnott (1667–1735) published in the *Philosophical Transactions of the Royal Society* a paper with the title, "An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes." In that paper Arbuthnott presented baptismal records from the years 1629 through 1710 giving the number of boys and girls baptized during those years. In each of the 82 years, without exception, the number of boys exceeded the number of girls by amounts varying from less than 3% in 1644 (4107 boys, 3997 girls) to more than 15% in 1659 (3209 boys, 2781 girls). Arbuthnott inferred correctly that the hypothesis that births of boys and girls were equally likely was not plausible, since it implied that an event with probability 2^{-82} had occurred. He even consulted a table of logarithms to write this number out in decimal form, so as to impress his readers:

$$\frac{1}{48360\,0000\,0000\,0000\,0000\,0000}$$

Exhibiting the usual haste to reach conclusions in such matters, Arbuthnott concluded that this constant imbalance must be the result of a divine plan to offset the higher mortality of males due to violence and accidents.¹⁵ He did not, for example, consider the possibility that more girls than boys were simply abandoned by mothers and fathers unable to support them. His final conclusion was that polygamy was against nature.

2.1. Quetelet. The first work on statistics proper was a treatise of 1835 entitled *Physique social*, written by the Belgian scientist Lambert Quetelet (1796–1874). Quetelet had been trained in both mathematics and astronomy, and he was familiar with the normal curve. He was the first to use it to describe variables other than those representing observational errors. He noticed certain analogies between probabilistic concepts and physical concepts, and he introduced them into social analysis. The most famous of these concepts was the "average person" (*l'homme moyen*), which he hoped could play a mathematical role similar to its physical analog, the center of gravity of a physical body.

2.2. Statistics in physics. One of the places in which individual phenomena are too numerous and too chaotic for analysis is in physics at the molecular level and below. Statistics has become an important tool in analyzing such systems. A very good example is thermodynamics, in which thermal energy is considered to be stored in a hypothetical (unobserved) translational and/or rotational motion of molecules against resisting forces that are equally hypothetical. In the simplest case, that of an ideal gas, there are no resisting forces and there is no rotational motion of molecules. All the thermal energy is stored as the translational kinetic energy of the molecules, which determine its temperature. At room temperature helium, which is monatomic, is the best approximation to an ideal gas.

In a way, thermodynamics, and in particular its famous second law, is only common sense, but physics needs to explain that common sense. Why can heat flow only from higher temperature to lower, just as water can flow only downhill? If temperature is determined by the translational kinetic energy of molecules, objects

¹⁵ In his day, death from contagious disease was at least as common in women as in men, perhaps even more common, due to the dangers of childbirth. Death from the debilities associated with old age was relatively uncommon.

at a higher temperature have molecules with higher kinetic energy—they are either more massive or moving faster. When two bodies are in contact, their molecules collide along the interface. Thermal energy then diffuses just as a gas diffuses when the boundaries confining it are removed. James Clerk Maxwell (1831–1879) created a theory of gases in which, he thought, the second law of thermodynamics could be violated. In an 1867 letter to Peter Guthrie Tait (1831–1901), he imagined a person or other agency, later dubbed “Maxwell’s demon” by Willam Thomson (Lord Kelvin, 1824–1907). The demon’s job was to stand guard at a small interface between two objects at different temperatures and allow only those molecules to pass through that would cause the temperature difference to increase.¹⁶ Statistically, it was *possible* that thermal energy might flow “uphill,” so to speak. The question was a quantitative one: How *likely* was that to happen?

The workings of this process can most easily be seen in the case of a sample of an ideal monatomic gas, whose pressure (P), volume (V), and absolute Kelvin temperature T satisfy the equation of state $PV = nRT = kT$, where n is the number of moles of gas present and R is a universal constant of proportionality. The quantity $S = k \ln(T^{3V/2})$ is called the *entropy* of the sample.¹⁷

The evolution of a thermally isolated system can be thought of as the effect of bringing many small samples of gas at different temperatures into contact. If a sample of n_1 moles of the ideal monatomic gas occupying volume V_1 at temperature T_1 is placed in thermally isolated contact with a sample of n_2 moles occupying volume V_2 at temperature T_2 , the total internal energy will be $\frac{3}{2}(k_1 + k_2)T = \frac{3}{2}k_1T_1 + \frac{3}{2}k_2T_2$, where T is the temperature after equilibrium is reached. Thus $T = (k_1T_1)/(k_1 + k_2) + (k_2T_2)/(k_1 + k_2) = cT_1 + (1 - c)T_2$. The ultimate entropy of the combined system will then be

$$(k_1 + k_2) \ln \left((cT_1 + (1 - c)T_2)^{\frac{3}{2}} (V_1 + V_2) \right).$$

This quantity is larger than the combined initial entropy of the two parts,

$$k_1 \ln (T_1^{3/2} V_1) + k_2 \ln (T_2^{3/2} V_2),$$

as one can see easily since $c = (k_1)/(k_1 + k_2)$.¹⁸ Thus, *entropy increases* for this system of two samples, and by extension in any thermally isolated system.

Maxwell began to urge a statistical view of thermodynamics in 1868, comparing the velocities of gas molecules with the white and black balls in the urn models that had been used for 150 years. In particular, he noted the tendency of these velocities to assume the normal distribution, as a consequence of the central limit theorem. When he became head of the Cavendish Laboratory at Cambridge in 1871, he said in his inaugural lecture that the statistical method

¹⁶ There is more to Maxwell’s demon than is implied here. Consider, for example, what energy is required for the demon to acquire the information about each molecule, decide whether to allow it to pass, and enforce the decision.

¹⁷ Strictly speaking, it is not possible to take the logarithm of a quantity having a physical dimension. The expression $\int (1/V) dV = \ln(V)$ should be interpreted in dimensionless terms. That is, V is really V/V_u , where V_u is a unit volume, and likewise $dV = 1/V_u dV$ and $\ln(V)$ is $\ln(V/V_u)$.

¹⁸ The function $3 \ln(x)/2$ is concave, so that a point on an arc of its graph lies above the chord of that arc, and (assuming without loss of generality that $V_1 \geq V_2$) $\ln(1 + (V_1)/(V_2)) \geq \ln((V_1)/(V_2)) \geq c \ln((V_1)/(V_2))$.

involves an abandonment of strict dynamical principles and an adoption of the mathematical methods belonging to the theory of probability. . . if the scientific doctrines most familiar to us had been those which must be expressed in this way, it is possible that we might have considered the existence of a certain kind of contingency a self-evident truth, and treated the doctrine of philosophical necessity as a mere sophism. [Quoted by Porter, (1986), pp. 201–202]

2.3. The metaphysics of probability and statistics. The statistical point of view required an adjustment in thinking. Maxwell appeared to like the indeterminacy that it introduced; Einstein was temperamentally opposed to it. Although deterministic and probabilistic models might both produce the same predictions because of the law of large numbers, there was a theoretical difference that could be seen clearly in thermodynamics. If the laws of Newtonian mechanics applied to the point-particles that theoretically made up, say, an ideal gas, the state of the gas should evolve equally well in either direction, since those mechanical laws are time-symmetric. Imagine, then, two identical containers containing the same number of molecules of an ideal gas that, at a given instant, are in exactly the same positions relative to the boundaries of the containers and such that each particle is moving with equal speed but in exactly opposite direction, to the particle in the corresponding place in the other container. By the laws of mechanics, the past states of each container must be the future states of the other. But then one of the two must be evolving in a direction that decreases entropy, in contradiction to the second law of thermodynamics.

The explanation of that “must be” is statistical. It is not *absolutely* impossible for the mechanical system to be in a state that would cause it to evolve, following the deterministic laws of mechanics, in a direction of decreasing entropy. But the initial conditions that lead to this evolution are *extremely unlikely*, so unlikely that no one ever expects to observe such a system. As an illustration, Newtonian mechanics can perfectly well explain water flowing uphill given that the initial velocities of all the water molecules are uphill. But no one ever expects these initial conditions to be satisfied in practice, in the absence of a tsunami. An additional consideration, which Maxwell regarded as relevant, was that in some cases initial-value problems do not have a unique solution. For example, the equation $\frac{3}{4} \cdot \frac{dy}{dt} - y^{1/4} = 0$ with initial condition $y = 0$ when $t = 0$ is satisfied for $t \geq 0$ by both relations $y \equiv 0$ and $y = t^{4/3}$. Shortly before his death, Maxwell wrote to Francis Galton:

There are certain cases in which a material system, when it comes to a phase in which the particular path which it is describing coincides with the envelope of all such paths may either continue in the particular path or take to the envelope (which in these cases is also a possible path) and which course it takes is not determined by the forces of the system (which are the same for both cases) but when the bifurcation of the path occurs, the system, ipso facto, invokes some determining principle which is extra physical (but not extra natural) to determine which of the two paths it is to follow. [Quoted in Porter, 1986, p. 206]

Statistics has been the focus of metaphysical debate, just like mathematics. For some early thinkers, such as Augustus de Morgan, the applications of probability were simply a matter of human ignorance: If we knew any reason for a system to be in one state rather than another, we would posit that reason as a physical law. In the absence of such a reason, all possible states are equally likely. A principle very close to this one is the basis of the second law of thermodynamics as now deduced in statistical physics. Yet other thinkers took a different point of view, positing some resemblance of the future to the past. This principle is the basis of the "frequentist" philosophy, which asserts that the probability of a future event is to be hypothesized from its occurrence in the past. The standard example is the question "What is the probability that the sun will rise tomorrow?" Assuming that we have adequate records that would have noted any exceptions to this very regular event over the past 5,000 years, a purely frequentist statistician would offer odds of 1,800,000 to 1 in favor of the event happening tomorrow. However, as William Feller (1906–1970) pointed out in his classic textbook of probability, our records really do not guarantee that there have been no exceptions. Our confidence that the sun rose in the remote past is based on the same considerations that give us confidence that it will rise tomorrow.

Opposed to the frequentists are the Bayesians, who believe it is possible to assign a probability to an event before a similar event has occurred. Classical probabilists, with their urn models, drawings from a deck of cards, and throws of dice, were in effect Bayesians who believed that symmetry considerations and intuition enabled people to assign probabilities to hypothetical events. The results of experiment, where available, helped to revise those assignments through Bayes' theorem: *If the events A_1, \dots, A_n are mutually exclusive and exhaustive, and B is any event, then*

$$P(A_k|B) = \frac{P(A_k \wedge B)}{P(B)} = \frac{P(A_k)P(B|A_k)}{\sum_j P(B|A_j)P(A_j)}.$$

The use of this formula is as follows. From some basic principles of symmetry, or purely subjectively, the events A_k are assigned hypothetical probabilities. Then the conditional probability of event B is computed assuming each of these events. After an experiment in which event B occurs, the probability of A_k is "updated" to the value of $P(A_k|B)$ computed from this formula. The simplest illustration is the case of a chest containing two drawers. Drawer 1 contains two gold coins, and drawer 2 contains a silver coin and a gold coin. The two events are A_1 : drawer 1 is chosen; A_2 : drawer 2 is chosen. Each is assigned a preliminary probability of $\frac{1}{2}$. Event B is "a gold coin is drawn from the drawer." The conditional probabilities are easily seen to be $P(B|A_1) = 1$, $P(B|A_2) = \frac{1}{2}$. If in fact a gold coin is drawn, then

$$P(A_1|B) = \frac{1 \cdot \frac{1}{2}}{1 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}} = \frac{\frac{1}{2}}{\frac{3}{4}} = \frac{2}{3}$$

and

$$P(A_2|B) = \frac{\frac{1}{2} \cdot \frac{1}{2}}{1 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

2.4. Correlations and statistical inference. The many intricate techniques that statisticians have developed today for analyzing data to determine if two variables are correlated are far too complex to be discussed in full here. But we cannot

leave this topic without at least mentioning one of the giants in this area, Karl Pearson (1856–1936), a British polymath who studied philosophy and law and was even called to the bar, although he never practiced. As professor at University College, London, he became interested in Darwin's theory of evolution and wrote a number of papers between 1893 and 1912 on the mathematics of evolution. In an 1893 paper he coined the term *standard deviation* to denote the natural unit of probability, and in 1900 he introduced the chi-square test of significance, a mainstay of applied statistics nowadays.¹⁹ Mathematically, the chi-square distribution with n degrees of freedom is the distribution of the sum of the squares of n independent standard normal distributions. What that means is that if the probability that X_k lies between a and b is given by the normal density,:

$$P(a \leq X_k \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}t^2} dt,$$

and each of these probabilities is independent of all the others, the probability that $X_1^2 + \cdots + X_n^2$ lies between 0 and c is given by the chi-square density with n degrees of freedom:

$$P(X_1^2 + \cdots + X_n^2 \leq c) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} \int_0^c x^{(n/2)-1} e^{-x/2} dx.$$

The chi-square distribution is useful because, if X_1, \dots, X_n are independent random variables with expected positive values μ_1, \dots, μ_n , the random variable

$$\chi^2 = \frac{(X_1 - \mu_1)^2}{\mu_1} + \cdots + \frac{(X_n - \mu_n)^2}{\mu_n}$$

has the chi-square distribution with $n - 1$ degrees of freedom. One can then determine whether actual deviations of the variables X_k from their expected values are likely to be random (hence whether *bias* is present) by computing the value of χ^2 and comparing it with a table of chi-square values.

To illustrate the connection between the chi-square and the standard normal distribution, Fig. 3 shows the frequency histogram for a computer experiment in which 1000 random values were computed for the sum of the squares of 10 standard normal random variables. This histogram is superimposed on the graph of the chi-square density function with 10 degrees of freedom.

The word *bias* in the preceding paragraph has a purely statistical meaning of “not random.” The rather pejorative meaning it has in everyday life is an indication of the connection that people tend to make between fairness and *equal outcomes*. If we find that some identifiable group of people is underrepresented or overrepresented in some other population—prisons or universities or other institutions—we proceed on the assumption that some cause is operating. In doing so, one must beware of jumping to conclusions, as Arbuthnott did. He was quite correct in his conclusion that the sexual imbalance was not a random deviation from a general rule of equality, but there are all kinds of possible explanations for the *bias*. An even larger sexual imbalance exists in China today, for example, as a result of the one-child policy of the Chinese government, combined with a traditional social pressure to produce male heirs. Evolutionary theory produces an explanation very similar to Arbuthnott's, but based on adaptation rather than intelligent, human-centered design.

¹⁹ The symbol χ^2 was Pearson's abbreviation for $x^2 + y^2$.

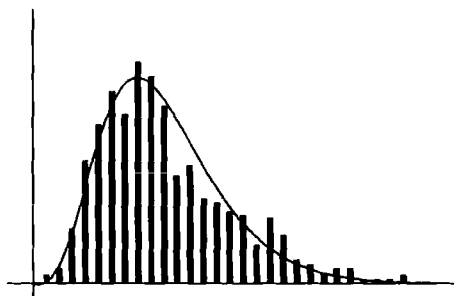


FIGURE 3. Comparison of the chi-square distribution and the frequencies for the sum of the squares of ten independent standard normal random variables when the experiment is performed 1000 times.

One of the many pitfalls of statistical inference was pointed out by Pearson's colleague George Udny Yule (1871- 1951) in 1903. Following up on Pearson's 1899 paper "On the spurious correlation produced by forming a mixture of heterogeneous but uncorrelated material," Yule produced a set of two 2×2 tables, each of which had no correlation, but produced a correlation when combined (see David and Edwards, 2001, p. 137). Yule's result was, for some reason, not given his name; but because it was publicized by Edward Hugh Simpson in 1951,²⁰ it came to be known as *Simpson's paradox*.²¹

Simpson's paradox is a counterintuitive oddity, not a contradiction. It arises frequently in practice. An example of it occurred in the admissions data from the graduate school of the University of California at Berkeley in 1973. These data raised some warning flags. Of the 12,763 applicants, 5232 were admitted, giving an admission rate of 41%. However, investigation revealed that 44% of the male applicants had been admitted and only 35% of the female applicants. There were 8442 male applicants, 3738 of whom were accepted, and 4321 female applicants, 1494 of whom were accepted. Simple chi-square testing showed that the hypothesis that these numbers represent a random deviation from a sex-independent acceptance rate of 41% was not plausible. There was unquestionably *bias*. The question was: Did this bias amount to discrimination? If so, who was doing the discriminating?

For more information on this case study and a very surprising conclusion, see "Sex bias in graduate admissions: data from Berkeley," *Science*, **187**, 7 February 1975, 398-404. In that paper, the authors analyzed the very evident *bias* in admissions to look for evidence of *discrimination*. Since admission decisions are made by the individual departments, it seemed logical to determine which departments had a noticeably higher admission rate for men than for women. Surprisingly, the authors found only four such departments (out of 101), and the imbalance resulting from those four departments was more than offset by six other departments that had a higher admission rate for women. It appears that the source of the bias was

²⁰ See "The interpretation of interaction in contingency tables," *Journal of the Royal Statistical Society*, Series B, **13**, 238-241.

²¹ The name *Simpson's paradox* goes back at least to the article of C. R. Blyth, "On Simpson's paradox and the sure-thing principle," in the *Journal of the American Statistical Association*, **67** (1972), 364-366.

hiding itself very well. The curious reader is referred to Problem 18.7, where the paradox is explained with an even more extreme example.

Questions and problems

18.1. Weather forecasters are evaluated for accuracy using the *Briers score*. The *a posteriori* probability of rain on a given day, judged from the observation of that day, is 0 if rain did not fall and 1 if rain did fall. A weather forecaster who said (the day before) that the chance of rain was 30% gets a Briers score of $30^2 = 900$ if no rain fell and $70^2 = 4900$ if rain fell. Imagine a very good forecaster, who over many years of observation learns that a certain weather pattern will bring rain 30% of the time. Also assume that for the sake of negotiating a contract that forecaster wishes to optimize (minimize) his or her Briers score. Should that forecaster state truthfully that the probability of rain is 30%? If we assume that the prediction and the outcome are independent events, we find that, for the days on which the true probability of rain is 30% the forecaster who makes a prediction of a 30% probability would in the long run average a Briers score of $0.3 \cdot 70^2 + 0.7 \cdot 30^2 = 2100$. This score is better (in the sense of a golf score—it is lower) than would result from randomly predicting a 100% probability of rain 30% of the time and a 0% probability 70% of the time. That strategy will be correct an expected 58% of the time ($.58 = .3^2 + .7^2$) and incorrect 42% of the time, resulting in a Briers score of $.42 \cdot 100^2 = 4200$. Let p be the actual probability of rain and x the forecast probability. Assuming that the event and the forecast are independent, show that the expected Briers score $10^4(p(1-x)^2 + (1-p)x^2)$ is minimized when $x = p$. [Note: If this result did not hold, a meteorologist who prized his/her reputation as a forecaster, based on the Briers measure, would be well advised to predict an incorrect probability, so as to get a better score for accuracy!]

18.2. We saw above that Cardano (probably) and Pascal and Leibniz (certainly) miscalculated some elementary probabilities. As an illustration of the counter-intuitive nature of many simple probabilities, consider the following hypothetical games. (A casino could probably be persuaded to open such games if there was enough public interest in them.) In game 1 the dealer lays down two randomly-chosen cards from a deck on the table and turns one face up. If that card is not an ace, no game is played. The cards are replaced in the deck, the deck is shuffled, and the game begins again. If the card is an ace, players are invited to bet against a fixed winning amount offered by the house that the other card is also an ace. What winning should the house offer (in order to break even in the long run) if players pay one dollar per bet?

In game 2 the rules are the same, except that the game is played only when the card turned up is the ace of hearts. What winning should the house offer in order to break even charging one dollar to bet? Why is this amount not the same as for game 1?

18.3. Use the Maclaurin series for $e^{-(1/2)t^2}$ to verify that the series given by de Moivre, which was

$$\sqrt{\frac{2}{\pi}} \left(\frac{1}{0! \cdot 1 \cdot 2} - \frac{1}{1! \cdot 3 \cdot 4} + \frac{1}{2! \cdot 5 \cdot 8} - \frac{1}{3! \cdot 7 \cdot 16} + \cdots \right),$$

represents the integral

$$\frac{1}{\sqrt{2\pi}} \int_0^1 e^{-\frac{1}{2}t^2} dt,$$

which is the area under a standard normal (bell-shaped) curve above the mean, but by at most one standard deviation, as given in many tables.

18.4. Use Daniel Bernoulli's concept of utility to explain why only a person with astronomical amounts of money should play a Petersburg paradox-type game. In your explanation, take account of what the utility of the stakes must be for a gambler versus the utility of the gain. Make an analogy between risk and work in this regard. A laborer exchanges time and effort for money; a gambler exchanges risk for potential gain. Remembering that all economic decisions are made "at the margin," at what point does additional work (or risk) not bring enough additional utility to be worth the exchange?

18.5. Radium-228 is an unstable isotope. Each atom of Ra-228 has a probability of 0.1145 (about 1 chance in 9, or about the probability of rolling a 5 with two dice) of decaying to form an atom of actinium within any given year. This means that the probability that the atom will survive the year as an atom of Ra-228 is $1 - 0.1145 = 0.8855$. Denote this "one-year survival" probability by p . Because any sample of reasonable size contains a huge number of atoms, that survival probability (0.8855) is the proportion of the weight of Ra-228 that we would expect to survive a year.

If you had one gram of Ra-228 to begin with, after one year you would expect to have $p = 0.8855$ grams. Each succeeding year, the weight of the Ra-228 left would be multiplied by p , so that after two years you would expect to have $p^2 = (0.8855)^2 = 0.7841$ grams. In general, after t years, if you started with W_0 grams, you would expect to have $W = W_0 p^t$ grams. Now push these considerations a little further and determine *how strongly* you can rely on this expectation. Recall Chebyshev's inequality, which says that the probability of being more than k standard deviations from the expected value is never larger than $(1/k)^2$. What we need to know to answer the question in this case is the standard deviation σ .

Our assumption is that each atom decays at random, independently of what happens to any other atom. This independence allows us to think that observing our sample for a year amounts to a large number of "independent trials," one for each atom. We test each atom to see if it survived as an Ra-228 atom or decayed into actinium. Let N_0 be the number of atoms that we started with. Assuming that we started with 1 gram of Ra-228, there will be $N_0 = 2.642 \cdot 10^{21}$ atoms of Ra-228 in the original sample.²² That is a very large number of atoms. The survival probability is $p = 0.8855$. For this kind of independent trial, as mentioned the standard deviation with N_0 trials is

$$\sqrt{N_0 p(1-p)} = \sqrt{\frac{p(1-p)}{N_0}} N_0.$$

We write the standard deviation in this odd-looking way so that we can express it as a fraction of the number N_0 that we started with. Since weights are proportional to the number of atoms, that same fraction will apply to the weights as well.

²² According to chemistry, the number of atoms in one gram of Ra-228 is the *Avogadro number* $6.023 \cdot 10^{23}$ divided by 228.

Put in the given values of p and N_0 to compute the fraction of the initial sample that constitutes one standard deviation. Since the original sample was assumed to be one gram, you can regard the answer as being expressed in grams. The use Chebyshev's inequality to estimate the probability that the amount of the sample remaining will differ from the theoretically predicted amount by 1 millionth of a gram (1 microgram, that is, 10^{-6} grams)? [Hint: How many standard deviations is one millionth of a gram?]

18.6. Analyze the revised probabilities in the problem of two drawers, one containing two gold coins, the other a gold and a silver coin, given an experiment in which event B occurs, if B is the event, "a silver coin is drawn."

18.7. Consider the case of 200 men and 200 women applying to a university consisting of only two different departments, and assume that the acceptance rates are given by the following table.

	Men	Women
Department A	120/160	32/40
Department B	8/40	40/160

Observe that the admission rate for men in department A is $\frac{3}{4}$, while that for women is $\frac{4}{5}$. In department B the admission rate for men is $\frac{1}{5}$ and for women it is $\frac{1}{4}$. In both cases, the people actually making the decisions are admitting a higher proportion of women than of men. Now explain the source of the bias, in our example and at Berkeley in simple, nonmathematical language.

CHAPTER 19

Logic and Set Theory

Logic has been an important part of western mathematics since the time of Plato. It also has a long history in other cultures, such as the Hindu and Buddhist culture (see Vidyabhusana, 1971). Logic became mathematized in the nineteenth century, in the work of mostly British mathematicians such as George Peacock (1791–1858), George Boole (1815–1864), William Stanley Jevons (1835–1882), and Augustus de Morgan, and a few Americans, notably Charles Sanders Peirce.

Set theory was the creation of nineteenth-century analysts and geometers, prominent among them Georg Cantor (1845–1918), whose inspiration came from geometry and analysis, mostly the latter. It resonated with the new abstraction that was entering mathematics from algebra and geometry, and its use by the French mathematicians Borel, Lebesgue, and Baire as the framework for their theories of integration and continuity helped to establish it as the foundation of all mathematics.

1. Logic

The mathematization of logic has a prehistory that goes back to Leibniz (not published in his lifetime), but we shall focus on mostly the nineteenth-century work. After a brief discussion of the preceding period, we examine the period from 1847 to 1930. This period opens with the treatises of Boole and de Morgan and closes with Gödel's famous incompleteness theorem. Our discussion is not purely about logic in the earlier parts, since the earlier writers considered both logical and probabilistic reasoning.

1.1. From algebra to logic. Leibniz was one of the first to conceive the idea of creating an artificial language in which to express propositions. He compared formal logic to the lines drawn in geometry as guides to thought. If the language encoded thought accurately, thought could be analyzed in a purely mechanical manner:

Then, in case of a difference of opinion, no discussion between two philosophers will be any longer necessary, as (it is not) between two calculators. It will rather be enough for them to take pen in hand, set themselves to the abacus and (if it so pleases, at the invitation of a friend) say to one another: *Let us calculate!* [Quoted by Bochenski, 1961, p. 275]

In another place he wrote:

Ordinary languages, though mostly helpful for the inferences of thought, are yet subject to countless ambiguities and cannot do the task of a calculus, which is to expose mistakes in inference. . . This remarkable advantage is afforded up to date only by the symbols of

arithmeticians and algebraists, for whom inference consists only in the use of characters, and a mistake in thought and in the calculus is identical. [Quoted by Bochenski, 1961, p. 275]

The ideal enunciated by Leibniz remains largely unfulfilled when it comes to settling philosophical disagreements. It reflects an oversimplified and optimistic view of human beings as basically rational creatures. This sort of optimism continued into the early nineteenth century, as exemplified by the *Handbook of Political Fallacies* by the philosopher Jeremy Bentham (1748–1832). But if the complex questions of the world of nature and society could not be mastered through logic alone, mathematics proved more amenable to the influences of logic. The influence, however, was bidirectional. In fact, there is a paradox, if one thinks of logic as being the rudder that steers mathematical arguments and keeps them from going astray. As Charles Sanders Peirce wrote in 1896, reviewing a book on logic:

It is a remarkable historical fact that there is a branch of science in which there has never been a prolonged dispute concerning the proper objects of that science. It is mathematics. . . Hence, we homely thinkers believe that, considering the immense amount of disputation there has always been concerning the doctrines of logic, and especially concerning those which would otherwise be applicable to settle disputes concerning the accuracy of reasonings in metaphysics, the safest way is to appeal for our logical principles to the science of mathematics. [Quoted in Bochenski, 1961, pp. 279–280]

Peirce seemed to believe that far from sorting out the mathematicians, logicians should turn to them for guidance. But we may dispute his assertion that there has never been a prolonged dispute about the proper objects of mathematics. Zeno confronted the Pythagoreans over that very question. In Peirce's own day, Kronecker and Cantor were at opposite ends of a dispute about what is and is not proper mathematics, and that discussion continues, politely, down to the present day. (See, for example, Hersh, 1997.)

Leibniz noted in the passage quoted above that algebra had the advantage of a precise symbolic language, which he held up as an ideal for clarity of communication. Algebra, in fact, was one of the mathematical sources of mathematical logic. When de Morgan translated a French algebra textbook into English in 1828, he defined algebra as “the part of mathematics in which symbols are employed to abridge and generalize the reasonings which occur in questions relating to numbers.” Thus, for de Morgan at the time, the symbols represented numbers, but *unspecified* numbers, so that reasoning about them applied to any particular numbers. Algebra was a ship anchored in numbers, but it was about to slip its anchor. In fact, only two years later (in 1830) George Peacock wrote a treatise on algebra in which he proposed that algebra be a purely symbolic science independent of any arithmetical interpretation. This step was a radical innovation at the time, considering that abstract groups, for example, were not to appear for several more decades. The assertion that the formula $(a - b)(a + b) = a^2 - b^2$ holds independently of any numerical values that replace a and b , for example, almost amounts to an axiomatic approach to mathematics. De Morgan's ideas on this subject matured during the 1830s, and at the end of the decade he wrote:

When we wish to give the idea of symbolical algebra... we ask, firstly, what symbols shall be used (without any reference to meaning); next, what shall be the laws under which such symbols are to be operated upon; the deduction of all subsequent consequences is again an application of common logic. Lastly, we explain the meanings which must be attached to the symbols, in order that they may have prototypes of which the assigned laws of operation are true. [Quoted by Richards, 1987, pp. 15–16]

This set of procedures is still the way in which mathematical logic operates, although the laws under which the symbols are to be operated on are now more abstract than de Morgan probably had in mind. To build a formal language, you first specify which sequences of symbols are to be considered “well-formed formulas,” that is, formulas capable of being true or false. The criterion for being well-formed must be purely formal, capable of being decided by a machine. Next, the sequences of well-formed formulas that are to be considered deductions are specified, again purely formally. The *syntax* of the language is specified by these two sets of rules, and the final piece of the construction, as de Morgan notes, is to specify its *semantics*, that is, the interpretation of its symbols and formulas. Here again, the modern world takes a more formal and abstract view of “interpretation” than de Morgan probably intended. For example, the semantics of propositional calculus consists of truth tables. After specifying the semantics, one can ask such questions as whether the language is consistent (incapable of proving a false proposition), complete (capable of proving all true propositions), or categorical (allowing only one interpretation, up to isomorphism).

In his 1847 treatise *Formal Logic*, de Morgan went further, arguing that “we have power to invent new meanings for all the forms of inference, in every way in which we have power to make new meanings of *is* and *is not* . . .” This focus on the meaning of *is* was very much to the point. One of the disputes that Peirce overlooked in the quotation just given is the question of what principles allow us to infer that an object “exists” in mathematics. We have seen this question in the eighteenth-century disagreement over what principles are allowed to define a function. In the case of symbolic algebra, where the symbols originally represented numbers, the existence question was still not settled to everyone’s liking in the early nineteenth century. That is why Gauss stated the fundamental theorem of algebra in terms of real factorizations alone. Here de Morgan was declaring the right to create mathematical entities by *fiat*, subject to certain restrictions. That enigmatic “exists” is indispensable in first-order logic, where the negation of “For every x , P ” is “For some x , not- P .” But what can “some” mean unless there actually *exist* objects x ? This defect was to be remedied by de Morgan’s friend George Boole.

In de Morgan’s formal logic, this “exists” remains hidden: When he talks about a class X , it necessarily has members. Without this assumption, even the very first example he gives is not a valid inference. He gives the following table by way of introduction to the symbolic logic that he is about to introduce:

<i>Instead of:</i>	<i>Write:</i>
All men will die	Every Y is X
All men are rational beings	Every Y is Z
Therefore some rational beings will die	Therefore some Z s are X ’s.

De Morgan's notation in this work was not the best, and very little of it has caught on. He used a parenthesis in roughly the same way as the modern notation for implication. For example, $X)Y$ denoted the proposition "Every X is a Y ." Nowadays we would write $X \supset Y$ (read " X horseshoe Y ") for " X implies Y ." The rest of his notation— $X : Y$ for "Some X 's are not Y s," $X.Y$ for "No X 's are Y s" and XY for "Some X 's are Y s"—is no longer used. For the negation of these properties he used lowercase letters, so that x denoted not- X . De Morgan introduced the useful "necessary" and "sufficient" language into implications: $X)Y$ meant that Y was *necessary* for X and X was *sufficient* for Y . He gave a table of the relations between X or x and Y or y for the relations $X)Y$, $X.Y$, $Y)X$, and $x.y$. For example, given that X implies Y , he noted that this relation made Y necessary for X , y an impossible condition for X , y a sufficient condition for x , and Y a contingent (not necessary, not sufficient, not impossible) condition for x .

For compound propositions, he wrote PQ for conjunction (his word), meaning both P and Q are asserted, and P, Q for disjunction (again, his word), meaning either P or Q . He then noted what are still known as *De Morgan's laws*:

The contrary of PQ is p, q . *Not both* is either not one or not the other, or not either. *Not either P nor Q* (which we might denote by $:P, Q$ or $.P, Q$) is logically '*not P and not Q* ' or pq : and this is then the contrary of P, Q .

De Morgan's theory of probability. De Morgan devoted three chapters (Chapters 9 through 11) to probability and induction, starting off with a very Cartesian principle:

That which we know, of which we are certain, of which we are well assured nothing could persuade us to the contrary, is the existence of our own minds, thoughts, and perceptions.

He then took the classical example of a certain proposition, namely that $2+2=4$ and showed by analyzing the meaning of 2, 4, and + that "It is true, no doubt, that 'two and two' is four, in amount, value, &c. but not in form, construction, definition, &c."¹ He continued:

There is no further use in drawing distinction between the knowledge which we have of our own existence, and that of two and two amounting to four. This absolute and inassailable feeling we shall call *certainty*. We have lower grades of knowledge, which we usually call *degrees of belief*, but they are really *degrees of knowledge*. . . It may seem a strange thing to treat *knowledge* as a magnitude, in the same manner as length, or weight, or surface. This is what all writers do who treat of probability. . . But it is not customary to make the statement so openly as I now do.

As this passage shows, for de Morgan probability was a subjective entity. He said that *degree of probability* meant *degree of belief*. In this way he placed himself firmly against the frequentist position, saying, "I throw away objective *probability*

¹ There is some intellectual sleight-of-hand here. The effectiveness of this argument depends on the reader's not knowing—and de Morgan's not stating—what is meant by addition of integers and by equality of integers, so that $2+2$ cannot be broken down into any terms simpler than itself. It really can be proved that $2+2=4$.

altogether, and consider the word as meaning the state of the mind with respect to an assertion, a coming event, or any other matter on which absolute knowledge does not exist." But subjectivity is not the same thing as arbitrariness. De Morgan, like us, would have labeled insane a person who asserted that the probability of rolling double sixes with a pair of dice is 50%. In fact, he gave the usual rules for dealing with probabilities of disjoint and independent events, and even stated Bayes' rule of inverse probability. He considered two urns, one containing six white balls and one black ball, the other containing two white balls and nine black ones. Given that one has drawn a white ball, he asked what the probability is that it came from the first urn. Noting that the probability of a white ball was $\frac{6}{7}$ in the first case and $\frac{2}{11}$ in the second, he concluded that the odds that it came from either of the two urns must be in the same proportion, $\frac{6}{7} : \frac{2}{11}$ or 33 : 7. He thus gave $\frac{33}{40}$ as the probability that the ball came from the first urn. This answer is in numerical agreement with the answer that would be obtained by Bayes' rule, but de Morgan did not think of it as revising a preliminary estimate of $\frac{1}{2}$ for the probability that the first urn was chosen.

1.2. Symbolic calculus. An example of the new freedom in the interpretation of symbols actually occurred somewhat earlier than the time of de Morgan, in Lagrange's algebraic approach to analysis. Thinking of Taylor's theorem as

$$\Delta_h f(x) = f(x+h) - f(x) = hDf(x) + \frac{1}{2!}h^2D^2f(x) + \frac{1}{3!}h^3D^3f(x) + \dots,$$

where $Df(x) = f'(x)$, and comparing with the Taylor's series of the exponential function,

$$e^t = 1 + t + \frac{1}{2!}t^2 + \frac{1}{3!}t^3 + \dots,$$

Lagrange arrived at the formal equation

$$\Delta_h = e^{hD} - 1.$$

Although the equation is purely formal and should perhaps be thought of only as a convenient way of remembering Taylor's theorem, it does suggest a converse relation

$$Df(x) = \frac{1}{h}(\ln(1 + \Delta_h))f(x) = \frac{1}{h}\left(\Delta_h f(x) + \frac{1}{2}\Delta_h^2 f(x) + \dots\right),$$

and this relation is literally true for polynomials $f(x)$. The formal use of this symbolic calculus may have been merely suggestive, but as Grattan-Guinness remarks (2000, p. 19), "some people regarded these methods as legitimate in themselves, not requiring foundations from elsewhere."

1.3. Boole's *Mathematical Analysis of Logic*. One such person was George Boole. In a frequently quoted passage from the introduction to his brief 1847 treatise *The Mathematical Analysis of Logic*, Boole wrote

[T]he validity of the processes of analysis does not depend upon the interpretation of the symbols which are employed but solely upon the laws of their combination. Every system of interpretation which does not affect the truth of the relations supposed is equally admissible, and it is thus that the same process may under one scheme of interpretation represent the solution of a question or the

properties of number, under another that of a geometrical problem, and under a third that of optics.

Here Boole, like de Morgan, was arguing for the freedom to create abstract systems and attach an interpretation to them later. This step was still something of an innovation at the time. It was generally accepted, for example, that irrational and imaginary numbers had a meaning in geometry but not in arithmetic. One could not, or should not, simply conjure them into existence. Cayley raised this objection shortly after the appearance of Boole's treatise (see Grattan-Guinness, 2000, p. 41), asking whether it made any sense to write $\frac{1}{2}x$. Boole replied by comparing the question to the existence of $\sqrt{-1}$, which he said was "a symbol (*i*) which satisfies particular laws, and especially this: $i^2 = -1$." In other words, when we are inventing a formal system, we are nearly omnipotent. Whatever we prescribe will hold for the system we define. If we want a square root of -1 to exist, it will exist (whatever "exist" may mean).

Logic and classes. Although set theory had different roots on the Continent, we can see its basic concept—membership in a class—in Boole's work. Departing from de Morgan's notation, he denoted a generic member of a class by an uppercase X , and used the lowercase x "operating on any subject," as he said, to denote the class itself. Then xy was to denote the class "whose members are both X 's and Y s." This language rather blurs the distinction between a set, its members, and the properties that determine what the members are; but we should expect that clarity would take some time to achieve. The connection between logic and set theory is an intimate one and one that is easy to explain. But the kind of set theory that logic alone would have generated was different from the geometric set theory of Georg Cantor, which is discussed in the next section.

The influence of the mathematical theory of probability on logic is both extensive and interesting. The subtitle of de Morgan's *Formal Logic is The Calculation of Inference, Necessary and Probable*, and, as noted above, three chapters (some 50 pages) of *Formal Logic* are devoted to probability and induction. Probability deals with events, whereas logic deals with propositions. The connection between the two was stated by Boole in his later treatise, *An Investigation of the Laws of Thought*, as follows:

[T]here is another form under which all questions in the theory of probabilities may be viewed; and this form consists in substituting for *events* the propositions which assert that those events have occurred, or will occur; and viewing the element of numerical probability as having reference to the *truth* of those *propositions*, not to the *occurrence* of the *events*.

Two events can combine in different ways: exactly one of E and F may occur, or E and F may both occur. If the events E and F are independent, the probability that both E and F occur is the *product* of their individual probabilities. If the two events cannot both occur, the probability that at least one occurs is the *sum* of their individual probabilities. More generally,

$$P(E \text{ or } F) + P(E \text{ and } F) = P(E) + P(F).$$

When these combinations of events are translated into logical terms, the result is a *logical calculus*.

The idea of probability 0 as indicating impossibility and probability 1 as indicating certainty must have had some influence on Boole's use of these symbols to denote "nothing" and "the universe." He expressed the proposition "all X 's are Y 's," for example, as $xy = x$ or $x(1 - y) = 0$. Notice that $1 - y$ appears, not $y - 1$, which would have made no sense. Here $1 - y$ corresponds to the things that are not- y . From there, it is not far to thinking of 0 as false and 1 as true. The difference between probability and logic here is that the probability of an event may be any number between 0 and 1, while propositions are either true or false.² These analogies were brought out fully in Boole's major work, to which we now turn.

1.4. Boole's *Laws of Thought*. Six years later, after much reflection on the symbolic logic that he and others had developed, Boole wrote an extended treatise, *An Investigation of the Laws of Thought*, which began by recapping what he had done earlier. The *Laws of Thought* began with a very general proposition that laid out the universe of symbols to be used. These were:

1st. Literal symbols, as x , y , &c., representing things as subjects of our conceptions.

2nd. Signs of operation, as $+$, $-$, \times , standing for those operations of the mind by which the conceptions of things are combined or resolved so as to form new conceptions involving the same elements.

3rd. The sign of identity, $=$.

And these symbols of Logic are in their use subject to definite laws, partly agreeing with and partly differing from the laws of the corresponding symbols in the science of Algebra.

Boole used $+$ to represent disjunction (or) and juxtaposition, used in algebra for multiplication, to represent conjunction (and). The sign $-$ was used to stand for "and not." In his examples, he used $+$ only when the properties were, as we would say, disjoint and $-$ only when the property subtracted was, as we would say, a subset of the property from which it was subtracted. He illustrated the equivalence of "European men and women" (where the adjective *European* is intended to apply to both nouns) with "European men and European women" as the equation $z(x+y) = zx + zy$. Similarly, to express the idea that the class of men who are non-Asiatic and white is the same as the class of white men who are not white Asiatic men, he wrote $z(x - y) = zx - zy$. He attached considerable importance to what he was later to call the *index law*, which expresses the fact that affirming a property twice conveys no more information than affirming it once. That is to say, $xx = x$, and he adopted the algebraic notation x^2 for xx . This piece of algebraization led him, by analogy with the rules $x0 = 0$ and $x1 = x$, to conclude that "the respective interpretations of the symbols 0 and 1 in the system of Logic are *Nothing* and *Universe*." From these considerations he deduced the principle of contradiction:

² Classical set theory deals with propositions of the form $x \in E$, which are either true or false: Either x belongs to E , or it does not, and there is no other possibility. The recently created *fuzzy set theory* restores the analogy with probability, allowing an element to belong partially to a given class and expressing the degree of membership by a function $\varphi(x)$ whose values are between 0 and 1. Thus, for example, whether a woman is pregnant or not is a classical set-theory question; whether she is tall or not is a fuzzy set-theory question. Fuzzy-set theorists point out that their subject is not subsumed by probability, since it deals with the properties of individuals, not those of large sets.

$x^2 = x \Rightarrow x(1 - x) = 0$, that is, no object can have a property and simultaneously not have that property.³

Boole was carried away by his algebraic analogies. Although he remained within the confines of his initial principles for a considerable distance, when he got to Chapter 5 he introduced the concept of *developing* a function. That is, for each algebraic expression $f(x)$, no matter how complicated, to find an equivalent linear expression $ax + b(1 - x)$, one that would have the same values as $f(x)$ for $x = 0$ and $x = 1$. That expression would obviously be $f(1)x + f(0)(1 - x)$. Boole gave a convoluted footnote to explain this simple fact by deriving it from Taylor's theorem and the idempotence property.

Like De Morgan, after discussing his 0-1 logic, Boole then turned to philosophy, metaphysics, and probability, placing himself in the philosophical camp of Poisson and de Morgan. He gave the now-familiar rule for the conditional probability of A given B as the probability of both A and B divided by the probability of B . He also gave a formal definition of independence, saying that two events are independent if "the probability of the happening of either of them is unaffected by our expectation of the occurrence or failure of the other." All of this was done in words, but could have been done symbolically, as he surely realized.

Application to jurisprudence. Nearly all of the early writers on probability, statistics, and logic had certain applications in mind, to insurance in the case of statistics, especially to the decisions of courts. The question of the believability of witnesses and the probability that a jury has been deceived interested Laplace, Quetelet, Poisson, de Morgan, and Boole, among others. Boole, for example, gave as an example, the following problem:

The probability that a witness A speaks the truth is p , the probability that another witness B speaks the truth is q , and the probability that they disagree in a statement is r . What is the probability that if they agree, their statement is true?

Boole gave the answer as $(p + q - r)/(2(1 - r))$. He claimed to prove as a theorem the following proposition:

From the records of the decisions of a court or deliberative assembly, it is not possible to deduce any definite conclusion respecting the correctness of the individual judgments of its members.

1.5. Venn. It is interesting to compare the mathematization of logic with the mathematization of probability. Both have ultimately been successful, but both were resisted to some extent as an intrusion of mathematics into areas of philosophy where it had no legitimate business. The case for removing mathematics from philosophy was made by John Venn, whose name is associated with a common tool of set theory: *Venn diagrams*, so-called, although the idea really goes back to Euler. In his book *The Logic of Chance*, which was first published in 1867, then revised a decade later and revised once again after another decade, Venn

³ Nowadays, a ring in which every element is idempotent, that is, the law $x^2 = x$ holds, is called a *Boolean ring*. It is an interesting exercise to show that such a ring is always commutative and of characteristic 2, that is, $x + x = 0$ for all x . The subsets of a given set form a Boolean ring when addition is interpreted as symmetric difference, that is, $A + B$ means "either A or B but not both."

proudly proclaimed that "Not only... will 'no knowledge of mathematics beyond the simple rules of Arithmetic' be required to understand these pages, but it is not intended that any such knowledge should be acquired by the process of reading them." Venn was particularly exercised about the attempts to apply probability theory in jurisprudence. Referring to Laplace, Quetelet, and the others, he wrote:

When they have searched for illustrations drawn from the practical business of life, they have very generally, but unfortunately, hit upon just the sort of instances which, as I shall endeavour to show hereafter, are among the very worst that could be chosen for the purpose. It is scarcely possible for any unprejudiced person to read what has been written about the credibility of witnesses by eminent writers, without his experiencing an invincible distrust of the principles which they adopt.

He went on to say that, although probability may require considerable mathematical knowledge, "the discussion of the fundamental principles on which the rules are based does not necessarily require any such qualification." Moreover,

The opinion that Probability, instead of being a branch of the general science of evidence which happens to make much use of mathematics, *is* a portion of mathematics, erroneous as it is, has yet been very disadvantageous to the science in several ways.

As one might expect, he took a dim view of the writings of de Morgan and Boole, saying that de Morgan had "given an investigation into the foundations of Probability as conceived by him, and nothing can be more complete and precise than his statement of principles and his deductions from them. If I could at all agree with these principles there would have been no necessity for the following essay." As for Boole, "Owing to his peculiar treatment of the subject, I have scarcely anywhere come into contact with any of his expressed opinions," a subtle, but acerbic way of saying that Boole had failed to convince anyone.

In Venn's view, expressed at the beginning of his fourth chapter, the practical application of probability in such matters as insurance was simply one more aspect of induction, the extrapolation of past experience into the future:

We cannot tell how many persons will be born or die in a year, or how many houses will be burnt or ships wrecked, without actually counting them. When we thus speak of "experience," we mean to employ the term in its widest signification; we mean experience supplemented by all the aids which inductive or deductive logic can afford. When, for instance, we have found the series which comprises the numbers of persons of any assigned class who die in successive years, we have no hesitation in extending it some way into the future as well as into the past. The justification of such a procedure must be sought in the ordinary canons of Induction.

Venn thus proclaimed himself a frequentist. The justification for applied probability and statistics was to be induction. But how firm a foundation was induction? The skeptical Scot David Hume (1711–1776) had leveled a devastating criticism

against the principles of induction and cause in the preceding century. Venn's reduction risked exposing probability theory to the same demolition.

1.6. Jevons. Both de Morgan and Boole used the syllogism or *modus ponens* (*inferring method*) as the basis of logical inference, although de Morgan did warn against an overemphasis on it. Their successor William Stanley Jevons, formulated this law algebraically and adjoined to it a principle of indirect inference, which amounted to inference by exhaustive enumeration of cases. The possibility of doing the latter by sorting through slips of paper led him to the conclusion that this sorting could be done by machine. Since he had removed much of the mathematical notation used by Boole, he speculated that the mathematics could be entirely removed from it. He also took the additional step of suggesting, rather hesitantly, that mathematics was itself a branch of logic. According to Grattan-Guinness (2000, p. 59), this speculation apparently had no influence on the mathematical philosophers who ultimately developed its implications, Russell and Frege.

2. Set theory

Set theory is all-pervasive in modern mathematics. It is the common language used to express concepts in all areas of mathematics. Because it is the language everyone writes in, it is difficult to imagine a time when mathematicians did not use the word *set* or think of sets of points. Yet that time was not long ago, less than 150 years. Before that time, mathematicians spoke of geometric figures. Or they spoke of points and numbers having certain properties, without thinking of those points and numbers as being assembled in a set. We have seen how concepts similar to those of set theory arose, in the notion of classes of objects having certain properties, in the British school of logicians. On the Continent, geometry and analysis provided the grounds for a development that resulted in a sort of "convergent evolution" with mathematical logic.

2.1. Technical background. Although the founder of set theory, Georg Cantor, was motivated by both geometry and analysis, for reasons of space we shall discuss only the analytic connection, which was the more immediate one. It is necessary to be slightly technical to explain how a problem in analysis leads to the general notion of a set and an ordinal number. We begin with the topic that Riemann developed for his 1854 lecture but did not use because Gauss preferred his geometric lecture. That topic was uniqueness of trigonometric series, and it was published in 1867, the year after Riemann's death. Riemann aimed at proving that if a trigonometric series converged to zero at every point, all of its coefficients were zero. That is,

$$\frac{1}{2}a_0 + \sum_{n=1}^{\infty}(a_n \cos nx + b_n \sin nx) \equiv 0 \implies a_n = 0 = b_n.$$

Riemann assumed that the coefficients a_n and b_n tend to zero, saying that it was clear to him that without that assumption, the series could converge only at isolated points.⁴ In order to prove this theorem, Riemann integrated twice to form the continuous function

$$F(x) = Ax + B + \frac{1}{4}a_0x^2 - \sum_{n=1}^{\infty} \frac{(a_n \cos nx + b_n \sin nx)}{n^2}.$$

⁴ Kronecker pointed out later that this assumption was dispensable; Cantor showed that it was deducible from the mere convergence of the series.

His object was to show that $F(x)$ must be a linear function, so that $G(x) = F(x) - Ax - B - \frac{1}{4}a_0x^2$ would be a quadratic polynomial that was also the sum of a uniformly convergent trigonometric series, and hence itself a constant, from which it would follow, first that $a_0 = 0$, and then that all the other a_n and all the b_n are zero. To that end, he showed that its generalized second derivative

$$F_g''(x) = \lim_{h \rightarrow 0} \frac{F(x+h) + F(x-h) - 2F(x)}{h^2}$$

was zero wherever the original series converged to zero.⁵ A weaker theorem, similar to the theorem that a differentiable function must be continuous, implied that

$$\lim_{h \rightarrow 0} \frac{(F(x+h) - F(x)) + (F(x-h) - F(x))}{h} = 0.$$

The important implication of this last result is that *the function $F(x)$ cannot have a corner*. If it has a right-hand derivative at a point, it also has a left-hand derivative at the point, and the two one-sided derivatives are equal. This fact, which at first sight appears to have nothing to do with set theory, was a key step in Cantor's work.

2.2. Cantor's work on trigonometric series. In 1872 Cantor published his first paper on uniqueness of trigonometric series, finishing the proof that Riemann had set out to give: that a trigonometric series that converges to zero at every point must have all its coefficients equal to zero. In following the program of proving that $F(x)$ is linear and hence constant, he observed that it was not necessary to assume that the series converged to zero at every point. A finite number of exceptions could be allowed, at which the series either diverged or converged to a nonzero value. For $F(x)$ is certainly continuous, and if it is linear on $[a, b]$ and also on $[b, c]$, the fact that it has no corners implies that it must be linear on $[a, c]$. Hence any isolated exceptional point b could be discounted.

The question therefore naturally arose: Can one allow an infinite number of exceptional points? Here one comes up against the Bolzano-Weierstrass theorem, which asserts that the exceptional points cannot all be isolated. They must have at least one point of accumulation. But exceptional points isolated from other exceptional points could be discounted, just as before. That left only their points of accumulation. If these were isolated—in particular, if there were only finitely many of them—the no-corners principle would once again imply uniqueness of the series.

Ordinal numbers. Cantor saw the obvious induction immediately. Denoting the set of points of accumulation of a set P (what we now call the derived set) by P' , he knew that $P' \supseteq P'' \supseteq P''' \supseteq \dots$. Thus, if at some finite level of accumulation points of accumulation points a finite set was obtained, the uniqueness theorem would remain valid. But the study of these sets of points of accumulation turned out to be even more interesting than trigonometric series themselves. No longer dealing with geometrically regular sets, Cantor was delving into point-set topology, as we now call it. No properties of a geometric nature were posited for the exceptional points he was considering, beyond the assumption that the sequence of derived sets must terminate at some finite level. Although the points of any particular set (as

⁵ Hermann Amandus Schwarz later showed that if $F_g''(x) \equiv 0$ on an open interval (a, b) , then $F(x)$ is linear on the closed interval $[a, b]$.

we now call it) might be easily describable, Cantor needed to discuss the general case. He needed the abstract concept of “sethood.” Cantor felt compelled to dig to the bottom of this matter and soon abandoned trigonometric series to write a series of papers on “infinite linear point-manifolds.”

Early on, he noticed the possibility of the transfinite. If the n th-level derived set is $P^{(n)}$, the nesting of these sets allows the natural definition of the derived set of infinite order $P^{(\infty)}$ as the intersection of all sets of finite order. But then one could consider derived sets even at the transfinite level: the derived set of $P^{(\infty)}$ could be defined as $P^{(\infty+1)} = (P^{(\infty)})'$. Cantor had discovered the infinite ordinal numbers. However, he did not at first recognize them as numbers, but rather regarded them as “symbols of infinity” (see Ferreirós, 1995).

Cardinal numbers. Cantor was not only an analyst, however. He had written his dissertation under Kronecker and Kummer on number-theoretic questions. Only two years after he wrote his first paper in trigonometric series, he noticed that his set-theoretic principles led to another interesting conclusion. The set of algebraic real or complex numbers is a countable set (as we would now say in the familiar language that we owe to Cantor), but the set of real numbers is not. Cantor had proved this point to his satisfaction in a series of exchanges of letters with Dedekind.⁶ Hence there must exist transcendental numbers. This second hierarchy of sets led to the concept of a cardinal number, two sets being of the same cardinality if they could be placed in one-to-one correspondence. To establish such correspondences, Cantor allowed himself certain powers of defining sets and functions that went beyond what mathematicians had been used to seeing. The result was a controversy that lasted some two decades.

Grattan-Guinness (2000, p. 125) has pointed out that Cantor emphasized five different aspects of point sets: their topology, dimension, measure, cardinality, and ordering. In the end, point-set topology was to become its own subject, and dimension theory became part of both algebraic and point-set topology. Measure theory became an important part of modern integration theory and had equally important applications to the theory of probability and random variables. Cardinality and ordering remained as an essential core of set theory, and the study of sets in relation to their complexity rather than their size became known as *descriptive set theory*.

Although descriptive set theory produces its own questions, it had at first a close relation to measure theory, since it was necessary to specify which sets could be measured. Borel was very careful about this procedure, allowing that the kinds of sets one could clearly define would have to be obtained by a finite sequence of operations, each of which was either a countable union or a countable intersection or a complementation, starting from ordinary open and closed sets. Ultimately those of a less constructive disposition than Borel honored him with the creation of

⁶ There are two versions of this proof, one due to Cantor and one due to Dedekind, but both involve getting nested sequences of closed intervals that exclude, one at a time, the elements of any given sequence $\{a_n\}$ of numbers. The intersection of the intervals must then contain a number not in the sequence. In his private speculations, Luzin noted that Cantor was actually assuming more than the mere *existence* of the countable set $\{a_n\}$. In order to construct a point not in it, one had to know something about each of its elements, enough to find a subinterval of the previous closed interval that would exclude the next element. On that basis, he concluded that Cantor had proved that there was no effective enumeration of the reals, not that the reals were uncountable. Luzin thus raised the question of what it could mean for an enumeration to “exist” if it was not effective. He too delved into philosophy to find out the meaning of “existence.”

the *Borel sets*, which is the smallest class that contains all closed subsets and also contains the complement of any of its sets and the union of any countable collection of its sets. This class can be “constructed” only by a transfinite induction.

Set theory, although it was an attempt to provide a foundation of clear and simple principles for all of mathematics, soon threw up its own unanswered mathematical questions. The most important was the continuum question. Cantor had shown that the set of all real numbers could be placed in one-to-one correspondence with the set of all subsets of the integers. Since he denoted the cardinality of the integers as \aleph_0 and the cardinality of the real numbers as \mathfrak{c} (where \mathfrak{c} stands for “continuum”), the question naturally arose whether there was any subset of the real numbers that had a cardinality between these two. Cantor struggled for a long time to settle this issue. One major theorem of set theory, known as the *Cantor-Bendixson theorem*,⁷ after Ivar Bendixson (1861–1935), asserts that every closed set is the union of a countable set and a perfect set, one equal to its derived set. Since it is easily proved that a nonempty perfect subset of the real numbers has cardinality \mathfrak{c} , it follows that every uncountable closed set contains a subset of cardinality \mathfrak{c} . Thus a set of real numbers having cardinality between \aleph_0 and \mathfrak{c} cannot be a closed set. Many mathematicians, especially the Moscow mathematicians after the arrival of Luzin as professor in 1915, worked on this problem. Luzin’s students Pavel Sergeevich Aleksandrov (1896–1982) and Mikhail Yakovlevich Suslin (1894–1919) proved that any uncountable Borel set must contain a nonempty perfect subset, and so must have cardinality \mathfrak{c} . Indeed, they proved this fact for a slightly larger class of sets called *analytic sets*. Luzin then proved that a set was a Borel set if and only if the set and its complement were both analytic sets.

The problem of the continuum remained open until 1938, when Kurt Gödel (1906–1978) partially closed it by showing that set theory is consistent with the continuum hypothesis and the axiom of choice,⁸ provided that it is consistent without them. Closure came to this question in 1963, when Paul Cohen (b. 1934—like Cantor, he began his career by studying uniqueness of trigonometric series representations) showed that the continuum hypothesis and the axiom of choice are independent of the other axioms of set theory.

2.3. The reception of set theory. If Venn believed that probability was an unwarranted intrusion of mathematics into philosophy, there were many mathematicians who believed that set theory was an equally unwarranted intrusion of philosophy into mathematics. One of those was Cantor’s teacher Leopold Kronecker. Although Cantor was willing to consider the existence of a transcendental number proved just because the real numbers were “too numerous” to be exhausted by the algebraic numbers, Kronecker preferred a more constructivist approach. His most famous utterance,⁹ and one of the most famous in all the history of mathematics, is: “The good Lord made the integers; everything else is a human creation.” (“*Die ganzen Zahlen hat der liebe Gott gemacht; alles andere ist Menschenwerk.*”) That is, the only infinity he admitted was the series of positive integers $1, 2, \dots$

⁷ Ferreirós (1995) points out that it was the desire to prove this theorem adequately, in 1882, that really led Cantor to treat transfinite ordinal numbers as numbers. He was helped toward this discovery by Dedekind’s pointing out to him the need to use finite ordinal numbers to define finite cardinal numbers.

⁸ Gödel actually included four additional assumptions in his consistency proof, one of the other two being that there exists a set that is analytic but is not a Borel set.

⁹ He made this statement at a meeting in Berlin in 1886 (see Grattan-Guinness, 2000, p. 122).

Beyond that point, everything was human-made and therefore had to be finite. If you spoke of a number or function, you had an obligation to say how it was defined. His 1845 dissertation, which he was unable to polish to his satisfaction until 1881, when he published it as “Foundations of an arithmetical theory of algebraic quantities” in honor of his teacher Kummer, shows how careful he was in his definitions. Instead of an arbitrary *field* defined axiomatically as we would now do, he wrote:

A domain of rationality is in general an *arbitrarily* bounded domain of magnitudes, but only to the extent that the concept permits. To be specific, since a domain of rationality can be enlarged only by the adjoining of arbitrarily chosen elements \mathfrak{A} , each arbitrary extension of its boundary requires the simultaneous inclusion of *all* quantities rationally expressible in terms of the new Element.

In this way, while one could enlarge a field to make an equation solvable, the individual elements of the larger field could still be described constructively. Kronecker’s concept of a general field can be described as “finitistic.” It is the minimal object that contains the necessary elements. Borel took this point of view in regard to measurable sets, and Hilbert was later to take a similar point of view in describing formal languages, saying that a meaningful formula must be obtained from a specified list of elements by a finite number of applications of certain rules of combination. This approach was safer and more explicit than, for example, Bernoulli’s original definition of a function as an expression formed “in some manner” from variables and constants. The “manner” was limited in a very definite way.

Cantor believed that Kronecker had delayed the publication of his first paper on infinite cardinal numbers. Whether that is the case or not, it is clear that Kronecker would not have approved of some of his principles of inference. As Grattan-Guinness points out, much of what is believed about the animosity between Cantor and Kronecker is based on Cantor’s own reports, which may be unreliable. Cantor was subject to periodic bouts of depression, probably caused by metabolic imbalances having nothing to do with his external circumstances. In fact, he had little to complain of in terms of the acceptance of his theories. It is true that there was some resistance to it, notably from Kronecker (until his death in 1891) and then from Poincaré. But there was also a great deal of support, from Weierstrass, Klein, Hilbert, and many others. In fact, as early as 1892, the journal *Bibliotheca mathematica* published a “Notice historique” on set theory by Giulio Vivanti (1859–1949), who noted that there had already been several expositions of the theory, and that it was still being developed by mathematicians, applied to the theory of functions of a real variable, and studied from a philosophical point of view.

2.4. Existence and the axiom of choice. In the early days Cantor’s set theory seemed to allow a remarkable amount of freedom in the “construction” or, rather, the calling into existence, of new sets. Cantor seems to have been influenced in his introduction of the term *set* by an essay that Dedekind began in 1872, but did not publish until 1887 (see Grattan-Guinness, 2000, p. 104), in which he referred to a “system” as “various things a, b, c, \dots comprehended from any cause under one point of view.” Dedekind defined a “thing” to be “any object of our thought.” Just as Descartes was able to conceive many things clearly and distinctly, mathematicians seemed to be able to form many “things” into “systems.” For example, given *any*

set A , one could conceive of another set whose members were the subsets of A . This set is nowadays denoted 2^A and called the *power set* of A . If A has a finite number n of elements, then 2^A has 2^n elements, counting the improper subsets \emptyset and A .

It was not long, however, before the indiscriminate use of this freedom to form sets led to paradoxes. The most famous of these is Russell's paradox, discussed in the next section. The source of the difficulty is that "existence" has a specialized mathematical meaning. The abstraction that comes with set theory has the consequence that much of the action in a proof takes place "offstage." That is, certain objects needed in a proof are called into existence by saying, "Let there be..." but no procedure for constructing them is given. Proofs relying on the abstract existence of such objects, when it is not possible to choose a particular object and examine it, became more and more common in the twentieth century. Indeed, much of measure theory, topology, and functional analysis would be impossible without such proofs. The principle behind these proofs later came to be known as *Zermelo's axiom*, after Ernst Zermelo (1871–1953), who first formulated it in 1904 to prove that every set could be well ordered.¹⁰ It was also known as the principle of free choice (in German, *Auswahlprinzip*) or, more commonly in English, the axiom of choice. In its broadest form this axiom states that *there exists a function f defined on the class of all nonempty sets such that $f(A) \in A$ for every nonempty set*. Intuitively, if A is nonempty, there exist elements of A , and $f(A)$ chooses one such element from every nonempty set.

This axiom is used in very many proofs, but probably the earliest (see Moore, 1982, p. 9) is Cantor's proof that a countable union of countable sets is countable. The proof goes as follows. Assume that A_1, A_2, \dots are countable sets, and let $A = A_1 \cup A_2 \cup \dots$. Then A is countable. For, let the sets A_j be enumerated, as follows

$$\begin{aligned} A_1 &= a_{11}, a_{12}, \dots, \\ A_2 &= a_{21}, a_{22}, \dots, \\ &\vdots \\ A_n &= a_{n1}, a_{n2}, \dots, \\ &\vdots \end{aligned}$$

Then the elements of A can be enumerated as follows: $a_{11}, a_{12}, a_{21}, a_{13}, a_{22}, a_{31}, \dots$, where the elements whose ranks are larger than the triangular number $T_n = n(n+1)/2$ but not larger than $T_{n+1} = (n+1)(n+2)/2$ are those for which the sum of the subscripts is $n+2$. There are $n+1$ such elements and $n+1$ such ranks. It is a very subtle point to notice that this proof assumes more than the mere existence of an enumeration of *each* of the sets, which is given in the hypothesis. It assumes the *simultaneous* existence of infinitely many enumerations, one for each set. The reasoning appears to be so natural that one would hardly question it. If a real choice exists at each stage of the proof, why can we not assume that infinitely many such choices have been made? As Moore notes, without the axiom of choice,

¹⁰ A set is *well ordered* if any two elements can be compared and every nonempty subset has a smallest element. The positive integers are well ordered by the usual ordering. The positive real numbers are not.

it is consistent to assume that the real numbers can be expressed as a countable union of countable sets.¹¹

Zermelo made this axiom explicit and showed its connection with ordinal numbers. The problem then was either to justify the axiom of choice, or to find a more intuitively acceptable substitute for it, or to find ways of doing without such "non-effective" concepts. A debate about this axiom took place in 1905 in the pages of the *Comptes rendus* of the French Academy of Sciences, which published a number of letters exchanged among Hadamard, Borel, Lebesgue, and Baire.¹² Borel had raised objections to Zermelo's proof that every set could be well-ordered on the grounds that it assumed an infinite number of enumerations. Hadamard thought it an important distinction that in some cases the enumerations were all independent, as in Cantor's proof above, but in others each depended for its definition on other enumerations having been made in correspondence with a smaller ordinal number. He agreed that the latter should not be used transfinitely. Borel had objected to using the axiom of choice nondenumeratively, but Hadamard thought that this usage brought no further damage, once a denumerable infinity of choices was allowed. He also mentioned the distinction due to Jules Tannery (1848–1910) between *describing* an object and *defining* it. To Hadamard, describing an object was a stronger requirement than defining it. To supply an example for him, we might mention a well-ordering of the real numbers, which is *defined* by the phrase itself, but effectively *indescribable*. Hadamard noted Borel's own work on analytic continuation and pointed out how it would change if the only power series admitted were those that could be effectively described. The difference, he said, belongs to psychology, not mathematics.

Hadamard received a response from Baire, who took an even more conservative position than Borel. He said that once an infinite set was spoken of, "the comparison, *conscious or unconscious*, with a bag of marbles passed from hand to hand must disappear completely."¹³ The heart of Baire's objection was Zermelo's *supposition* that to each (nonempty) subset of a set M there corresponds one of its elements." As Baire said, "all that it proves, as far as I am concerned, is that we do not perceive a contradiction" in imagining any set well-ordered.

Responding to Borel's request for his opinion, Lebesgue gave it. As far as he was concerned, Zermelo had very ingeniously shown how to solve problem A (to well-order any set) provided one could solve problem B (to choose an element from every nonempty subset of a given set). He remarked, probably with some irony, that, "Unfortunately, problem B is not easy to resolve, it seems, except for the sets that we know how to well-order." Lebesgue mentioned a concept that was to play a large role in debates over set theory, that of "effectiveness," roughly what we would call constructibility. He interpreted Zermelo's claim as the assertion that a well-ordering *exists* (that word again!) and asked a question, which he said was "hardly new": *Can one prove the existence of a mathematical object without defining it?* One would think not, although Zermelo had apparently proved the existence of a well-ordering (and Cantor had proved the existence of a transcendental number) without *describing* it. Lebesgue and Borel preferred the verb *to name* (*nommer*)

¹¹ Not *every* countable union of countable sets is uncountable, however; the rational numbers remain countable, because an explicit counting function can be constructed.

¹² These letters were translated into English and published by Moore (1982, pp. 311–320).

¹³ Luzin said essentially the same in his journal: What makes the axiom of choice seem reasonable is the picture of reaching into a set and helping yourself to an element of it.

when referring to an object that was defined effectively, through a finite number of uses of well-defined operations on a given set of primitive objects.

After reading Lebesgue's opinion, Hadamard was sure that the essential distinction was between what is determined and what is described. He compared the situation with the earlier debate over the allowable definitions of a function. But, he said, uniqueness was not an issue. If one could say "For each x , there exists a number satisfying. . . . Let y be this number," surely one could also say "For each x , there exists an infinity of numbers satisfying. . . . Let y be one of these numbers." But he put his finger squarely on one of the paradoxes of set theory (the Burali-Forti paradox, discussed in the next section). "It is the very existence of the set W that leads to a contradiction. . . the general definition of the word *set* is incorrectly applied." (Question to ponder: What is the definition of the word *set*?)

The validity and value of the axiom of choice remained a puzzle for some time. It leads to short proofs of many theorems whose statements are constructive. For example, it proves the existence of a nonzero translation-invariant Borel measure on any locally compact Abelian group. Since such a measure is provably unique (up to a constant multiple), there ought to be effective proofs of its existence that do not use the axiom of choice (and indeed there are). One benefit of the 1905 debate was a clarification of equivalent forms of the axiom of choice and an increased awareness of the many places where it was being used. A list of important theorems whose proof used the axiom was compiled for Luzin's seminar in Moscow in 1918. The list showed, as Luzin wrote in his journal, that "almost nothing is proved without it." Luzin was horrified, and spent some restless nights pondering the situation.

The axiom of choice is ubiquitous in modern analysis; almost none of functional analysis or point-set topology would remain if it were omitted entirely (although weaker assumptions might suffice). It is fortunate, therefore, that its consistency with, and independence of, the other axioms of set theory has been proved. However, the consequences of this axiom are suspiciously strong. In 1924 Alfred Tarski (1901–1983) and Stefan Banach (1892–1945) deduced from it that any two sets A and B in ordinary three-dimensional Euclidean space, each of which contains some ball, can be decomposed into pairwise congruent subsets. This means, for example, that a cube the size of a grain of salt (set A) and a ball the size of the Sun (set B) can be written as disjoint unions of sets A_1, \dots, A_n and B_1, \dots, B_n respectively such that A_i is congruent to B_i for each i . This result (the Banach–Tarski paradox) is very difficult to accept. It can be rationalized only by realizing that the notion of existence in mathematics has no metaphysical content. To say that the subsets A_i , B_i "exist" means only that a certain formal statement beginning $\exists \dots$ is deducible from the axioms of set theory.

2.5. Doubts about set theory. The powerful and counterintuitive results obtained from the axiom of choice naturally led to doubts about the consistency of set theory. Since it was being inserted under the rest of mathematics as a foundation, the consistency question became an important one. A related question was that of completeness. Could one provide a foundation for mathematics, that is, a set of basic objects and rules of proof, that would allow any meaningful proposition to be proved true or false? The two desirable qualities are in the abstract opposed to each other, just as avoiding disasters and avoiding false alarms are opposing goals.

The most influential figure in mathematical logic during the twentieth century was Gödel. The problems connected with consistency and completeness of arithmetic, the axiom of choice, and many others all received a fully satisfying treatment at his hands that settled many old questions and opened up new areas of investigation. In 1931, he astounded the mathematical world by producing a proof that any consistent formal language in which arithmetic can be encoded is necessarily incomplete, that is, contains statements that are true according to its metalanguage but not deducible within the language itself. The intuitive idea behind the proof is a simple one, based on the statement that follows:

This statement cannot be proved.

Assuming that this statement has a meaning—that is, its context is properly restricted so that “proved” has a definite meaning—we can ask whether it is *true*. The answer must be positive if the system in which it is made is consistent. For if this statement is false, by its own content, it *can* be proved; and in a consistent deductive system, a false statement cannot be proved. Hence we agree that the statement is true, but, again by its own content, it cannot be proved.

The example just given is really nonsensical, since we have not carefully delineated the universe of axioms and rules of inference in which the statement is made. The word “proved” that it contains is not really defined. Gödel, however, took an accepted formalization of the axioms and rules of inference for arithmetic and showed that the metalanguage of arithmetic could be encoded within arithmetic. In particular each formula can be numbered uniquely, and the statement that formula n is (or is not) deducible from those rules can itself be coded as a well-formed formula of arithmetic. Then, when n is chosen so that the statement, “Formula number n cannot be proved” happens to *be* formula n , we have exactly the situation just described. Gödel showed how to construct such an n . Thus, if Gödel’s version of arithmetic is consistent, it contains statements that are formally undecidable; that is, they are true (based on the metalanguage) but not deducible. This is Gödel’s first incompleteness theorem. His second incompleteness theorem is even more interesting: *The assertion that arithmetic is consistent is one of the formally undecidable statements.*¹⁴ If the formalized version of arithmetic that Gödel considered is consistent, it is incapable of proving itself so. It is doubtful, however, that one could truly formalize every kind of argument that a rational person might produce. For that reason, care should be exercised in drawing inferences from Gödel’s work to the actual practice of mathematics.

3. Philosophies of mathematics

Besides Cantor, other mathematicians were also considering ways of deriving mathematics logically from simplest principles. Gottlob Frege (1848–1925), a professor in Jena, who occasionally lectured on logic, attempted to establish logic on the basis of “concepts” and “relations” to which were attached the labels *true* or *false*. He was the first to establish a complete predicate calculus, and in 1884 wrote a treatise called *Grundgesetze der Arithmetik* (*Principles of Arithmetic*). Meanwhile in Italy, Giuseppe Peano (1858–1939) was axiomatizing the natural numbers. Peano took the successor relation as fundamental and based his construction of the natural

¹⁴ Detlefsen (2001) has analyzed the meaning of proving consistency in great detail and concluded that the generally held view of this theorem—that the consistency of a “sufficiently rich” theory cannot be proved by a “finitary” theory—is incorrect.

numbers on this one relation and nine axioms, together with a symbolic logic that he had developed. The work of Cantor, Frege, and Peano attracted the notice of a young student at Cambridge, Bertrand Russell, who had written his thesis on the philosophy of Leibniz. Russell saw in this work confirmation that mathematics is merely a prolongation of formal logic. This view, that mathematics can be deduced from logic without any new axioms or rules of inference, is now called *logicism*. Gödel's work was partly inspired by it, and can be interpreted as a counterargument to its basic thesis—that mathematics can be axiomatized. Logicism had encountered difficulties still earlier, however. Even the seemingly primitive notion of membership in a set turned out to require certain caveats.

3.1. Paradoxes. In 1897 Peano's assistant Cesare Burali-Forti (1861–1931), apparently unintentionally, revealed a flaw in the ordinal numbers.¹⁵ To state the problem in the clear light of hindsight, if two ordinal numbers satisfy $x < y$, then $x \in y$, but $y \notin x$. In that case, what are we to make of the set of all ordinal numbers? Call this set A . Like any other ordinal number, it has a successor $A + 1$ and $A \in A + 1$. But since $A + 1$ is an ordinal number, we must also have $A + 1 \in A$, and hence $A < A + 1$ and $A + 1 < A$. This was the first paradox of uncritical set theory, but others were to follow.

The most famous paradox of set theory arose in connection with cardinal numbers rather than ordinal numbers. Cantor had defined equality between cardinal numbers as the existence of a one-to-one correspondence between sets representing the cardinal numbers. Set B has larger cardinality than set A if there is no function $f : A \rightarrow B$ that is “onto,” that is, such that every element of B is $f(x)$ for some $x \in A$. Cantor showed that the set of all subsets of A , which we denote 2^A , is always of larger cardinality than A , so that there can be no largest cardinal number. If $f : A \rightarrow 2^A$, the set $C = \{t \in A : t \notin f(t)\}$ is a subset of A , hence an element of 2^A , and it cannot be $f(x)$ for any $x \in A$. For if $C = f(x)$, we ask whether $x \in C$ or not. If $x \in C$, then $x \in f(x)$ and so by definition of C , $x \notin C$. On the other hand, if $x \notin C$, then $x \notin f(x)$, and again by definition of C , $x \in C$. Since the whole paradox results from the assumption that $C = f(x)$ for some x , it follows that no such x exists, that is, the mapping f is not “onto.” This argument was at first disputed by Russell, who wrote in an essay entitled “Recent work in the philosophy of mathematics” (1901) that “the master has been guilty of a very subtle fallacy.” Russell thought that there was a largest set, the set of *all* sets. In a later reprint of the article he added a footnote explaining that Cantor was right.¹⁶ Russell's first attempt at a systematic exposition of mathematics as he thought it ought to be was his 1903 work *Principles of Mathematics*. According to Grattan-Guinness (2000, p. 311), Russell removed his objection to Cantor's proof and published his paradox in this work, but kept the manuscript of an earlier version, made before he was able to work out where the difficulty lay.

To explain Russell's paradox, consider the set of all sets. We must, by its definition, believe it to be *equal* to the set of all its subsets. Therefore the mapping $f(E) = E$ should have the property that Cantor says no mapping can have. Now

¹⁵ Moore (1982, p. 59) notes that Burali-Forti himself did not see any paradox and (p. 53) that the difficulty was known earlier to Cantor.

¹⁶ Moore (1982, p. 89) points out that Zermelo had discovered Russell's paradox two years before Russell discovered it and had written to Hilbert about it. Zermelo, however, did not consider it a very troubling paradox. To him it meant only that no set should contain all of its subsets as elements.

if we apply Cantor's argument to this mapping, we are led to consider $S = \{E : E \notin E\}$. By definition of the mapping f we should have $f(S) = S$, and so, just as in the case of Cantor's argument, we ask if $S \in S$. Either way, we are led to a contradiction. This result is known as *Russell's paradox*.

After Russell had straightened out the paradox with a theory of types, he collaborated with Alfred North Whitehead on a monumental derivation of mathematics from logic, published in 1910 as *Principia mathematica*.

3.2. Formalism. A different view of the foundations of mathematics was advanced by Hilbert, who was interested in the problem of axiomatization (the axiomatization of probability theory was the sixth of his famous 23 problems) and particularly interested in preserving as much as possible of the freedom to reason that Cantor had provided while avoiding the uncomfortable paradoxes of logicism. The essence of this position, now known as formalism, is the idea stated by de Morgan and Boole that the legal manipulation of the symbols of mathematics and their interpretation are separate issues. Hilbert is famously quoted as having claimed that the words *point*, *line*, and *plane* should be replaceable by *table*, *chair*, and *beer mug* when a theorem is stated. Grattan-Guinness (2000, p. 208) notes that Hilbert may not have intended this statement in quite the way it is generally perceived and may not have thought the matter through at the time. He also notes (p. 471) that Hilbert never used the name *formalism*. Characteristic of the formalist view is the assumption that any mathematical object whatever may be defined, provided only that the definition does not lead to a contradiction. Cantor was a formalist in this sense (Grattan-Guinness, 2000, p. 119). In the formalist view mathematics is the study of formal systems, but the rules governing those systems must be stated with some care. In that respect, formalism shares some of the caution of the earlier constructivist approach. It involves a strict separation between the symbols and formulas of mathematics and the meaning attached to them, that is, a distinction between syntax and semantics. Hilbert had been interested in logical questions in the 1890s and early 1900s, but his work on formal languages such as propositional calculus dates from 1917. In 1922, when the intuitionists (discussed below) were publishing their criticism of mathematical methodologies, he formulated his own version of mathematical logic. In it he introduced the concept of metamathematics, the study whose subject matter is the structure of a mathematical system.¹⁷ A formal language consists of certain rules for recognizing legitimate formulas, certain formulas called axioms, and certain rules of inference (such as syllogism, generalization over unspecified variables, and the rules for manipulating equations). These elements make up the syntax of the language. One can therefore always tell by following clearly prescribed rules whether a formula is meaningful (well formed) and whether a sequence of formulas constitutes a valid deduction. To avoid infinity in this system while preserving sufficient generality, Hilbert resorted to a "finitistic" device called a *schema*. Certain basic formulas are declared to be legitimate by fiat. Then a few rules are adopted, such as the rule that if A and B are legitimate formulas, so is $[A \Rightarrow B]$. This way of defining legitimate (well-formed) formulas makes it possible to determine in a finite number of steps whether or not a formula is well formed. It replaces the synthetic constructivist approach with an analytic

¹⁷ This distinction had been introduced by L. E. J. Brouwer in his 1907 thesis, but not given a name and never developed (see Grattan-Guinness, 2000, p. 481).

approach (which can be reversed, once the analysis is finished, to synthesize a given well-formed formula from primitive elements).

The formalist approach makes a distinction between statements *of* arithmetic and statements *about* arithmetic. For example, the assertion that there are no positive integers x, y, z such that $x^3 + y^3 = z^3$ is a statement *of* arithmetic. The assertion that this statement can be proved from the axioms of arithmetic is a statement *about* arithmetic. The *metalanguage*, in which statements are made about arithmetic, contains all the meaning to be assigned to the propositions of arithmetic. In particular, it becomes possible to distinguish between what is true (that is, what can be known to be true from the metalanguage) and what is provable (what can be deduced within the object language). Two questions thus arise in the metalanguage: (1) *Is every deducible proposition true?* (the problem of consistency); (2) *Is every true proposition deducible?* (the problem of completeness). As we saw in Section 2, Gödel showed that the answer, for first-order recursive arithmetic and more generally for systems of that type, is very pessimistic. This language is not complete and is incapable of proving its own consistency.

3.3. Intuitionism. The most cautious approach to the foundations of mathematics, known as *intuitionism*, was championed by the Dutch mathematician Luitzen Egbertus Jan Brouwer (1881–1966). Brouwer was one of the most mystical of mathematicians, and his mysticism crept into his early work. He even published a pamphlet in 1905, claiming that true happiness came from the inner world, and that contact with the outer world brought pain (Franchella, 1995, p. 305). In his dissertation at the University of Amsterdam in 1907, he criticized the logicism of Russell and Zermelo's axiom of choice. Although he was willing to grant the validity of constructing each particular denumerable ordinal number, he questioned whether one could meaningfully form a set of all denumerable ordinals.¹⁸ In a series of articles published from 1918 to 1928, Brouwer laid down the principles of intuitionism. These principles include the rejection not only of the axiom of choice beyond the countable case, but also of proof by contradiction. That is, the implication "A implies not-(not-A)" is accepted, but not its converse, "Not-(not-A) implies A." Intuitionists reject any proof whose implementation leaves choices to be made by the reader. Thus it is not enough in an intuitionist proof to say that objects of a certain kind exist. One must choose such an object and use it for the remainder of the proof. This extreme caution has rather drastic consequences. For example, the function $f(x)$ defined in ordinary language as

$$f(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

is not considered to be defined by the intuitionists, since there are ways of defining numbers x that do not make it possible to determine whether the number is negative or positive. For example, is the number $(-1)^n$, where n is the trillionth decimal digit of π , positive or negative? This restrictedness has certain advantages, however. The objects that are acceptable to the intuitionists tend to have pleasant properties. For example, every rational-valued function of a rational variable is continuous.

The intuitionist rejection of proof by contradiction needs to be looked at in more detail. Proof by contradiction was always used somewhat reluctantly by

¹⁸ This objection seems strange at first, but the question of whether an effectively defined set must have effectively defined members is not at all trivial.

mathematicians, since such proofs seldom give insight into the structures being studied. For example, Euclid's proof that there are infinitely many primes proceeds by assuming that the set of prime numbers is a finite set $P = \{p_1, p_2, \dots, p_n\}$ and showing that in this case the number $1 + p_1 \cdots p_n$ must either itself be a prime number or be divisible by a prime different from p_1, \dots, p_n , which contradicts the original assumption that p_1, \dots, p_n formed the entire set of prime numbers.

The appearance of starting with a false assumption and deriving a contradiction can be avoided here by stating the theorem as follows: If there exists a set of n primes p_1, \dots, p_n , there exists a set of $n + 1$ primes. The proof is exactly as before. Nevertheless, the proof is still not intuitionistically valid, since there is no way of saying whether or not $1 + p_1 \cdots p_n$ is prime.

In 1928 and 1929, a quarter-century after the debate over Zermelo's axiom of choice, there was debate about intuitionism in the bulletin of the Belgian Royal Academy of Sciences. Two Belgian mathematicians, M. Barzin and A. Errera, had argued that Brouwer's logic amounted to a three-valued logic, since a statement could be true, false, or undecidable. The opposite point of view was defended by two distinguished Russian mathematicians, Aleksandr Yakovlevich Khinchin (1894–1959) and Valerii Ivanovich Glivenko (1897–1940). Barzin and Errera had suggested that to avoid three-valued logic, intuitionists ought to adopt as an axiom that if p implies " q or r ", then either p implies q or p implies r ,¹⁹ and also that if " p or q " implies r , then p implies r and q implies r . Starting from these principles of Barzin and Errera and the trivial axiom " p or not- p " implies " p or not- p ", Khinchin deduced that p implies not- p and not- p implies p , thus reducing the suggestions of Barzin and Errera to nonsense. Glivenko took only a little longer to show that, in fact, Brouwer's logic was not three-valued. He proved that the statement " p or not- p is false" is false in Brouwer's logic, and ultimately derived the theorem that the statement " p is neither true nor false" is false (see Novosyolov, 2000).

A more "intuitive" objection to intuitionism is that intuition by its nature cannot be codified as a set of rules. In adopting such rules, the intuitionists were not being intuitionistic in the ordinary sense of the word. In any case, intuitionist mathematics is obviously going to be somewhat sparser in results than mathematics constructed on more liberal principles. That may be why it has attracted only a limited group of adherents.

3.4. Mathematical practice. The paradoxes of naive set theory (such as Russell's paradox) were found to be avoidable if the word *class* is used loosely, as Cantor had previously used the word *set*, but the word *set* is restricted to mean only a class that is a member of some other class. (Classes that are not sets are called *proper classes*.) Then to belong to a class A , a class B must not only fulfill the requirements of the definition of the class A but must also be known in advance to belong to some (possibly different) class.

This approach avoids Russell's paradox. The class $C = \{x : x \notin x\}$ is a class; its elements are those classes that *belong to some class and* are not elements of themselves. If we now ask the question that led to Russell's paradox—Is C a member of itself?—we do not reach a contradiction. If we assume $C \in C$, then we conclude that $C \notin C$, so that this assumption is not tenable. However, the opposite assumption, that $C \notin C$, is acceptable. It no longer leads to the conclusion

¹⁹ In the currently accepted semantics (metalanguage) of intuitionistic propositional calculus, if " q or r " is a theorem, then either q is a theorem or r is a theorem.

that $C \in C$. For an object x to belong to C , it no longer suffices that $x \notin x$; it must also be true that $x \in A$ for some class A , an assumption not made for the case when x is C . A complete set of axioms for set theory avoiding all known paradoxes was worked out by Paul Bernays (1888–1977) and Adolf Fraenkel (1891–1965). It forms part of the basic education of mathematicians today. It is generally accepted because mathematics can be deduced from it. However, it is very far from what Cantor had hoped to create: a clear, concise, and therefore *obviously* consistent foundation for mathematics. The axioms of set theory are extremely complicated and nonintuitive, and far less obvious than many things deduced from them. Moreover, their consistency is not only not obvious, it is even unprovable. In fact, one textbook of set theory, *Introduction to Set Theory*, by J. Donald Monk (McGraw-Hill, New York, 1969), p. 22, asserts of these axioms: “Naturally no inconsistency has been found, and *we have faith* that the axioms are, in fact, consistent”! (Emphasis added.)

Questions and problems

19.1. Bertrand Russell pointed out that some applications of the axiom of choice are easier to avoid than others. For instance, given an infinite collection of pairs of shoes, describe a way of choosing one shoe from each pair. Could you do the same for an infinite set of pairs of socks?

19.2. Prove that $C = \{x : x \notin x\}$ is a proper class, not a set, that is, it is not an element of any class.

19.3. Suppose that the only allowable way of forming new formulas from old ones is to connect them by an implication sign; that is, given that A and B are well formed, $[A \Rightarrow B]$ is well formed, and conversely, if A and B are not both well formed, then neither is $[A \Rightarrow B]$. Suppose also that the only basic well-formed formulas are p , q , and r . Show that

$$[[p \Rightarrow r] \Rightarrow [[p \Rightarrow q] \Rightarrow r]]$$

is well formed but

$$[[p \Rightarrow r] \Rightarrow [r \Rightarrow]]$$

is not. Describe a general algorithm for determining whether a finite sequence of symbols is well formed.

19.4. Consider the following theorem. There exists an irrational number that becomes rational when raised to an irrational power. *Proof:* Consider the number $\theta = \sqrt{3}^{\sqrt{2}}$. If this number is rational, we have an example of such a number. If it is irrational, the equation $\theta^{\sqrt{2}} = \sqrt{3}^2 = 3$ provides an example of such a number. Is this proof intuitionistically valid?

19.5. Show that any two distinct *Fermat numbers* $2^{2^m} + 1$ and $2^{2^n} + 1$, $m < n$, are relatively prime. (Use mathematical induction on n .) Apply this result to deduce that there are infinitely many primes. Would this proof of the infinitude of the primes be considered valid by an intuitionist?

19.6. Suppose that you prove a theorem by assuming that it is false and deriving a contradiction. What you have then proved is that either the axioms you started with are inconsistent or the assumption that the theorem is false is itself false.

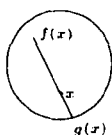


FIGURE 1. The Brouwer fixed-point theorem.

Why should you conclude the latter rather than the former? Is this why some mathematicians have claimed that the practice of mathematics requires faith?

19.7. What are the advantages, if any, of building a theory by starting with abstract definitions, then later proving a structure theorem showing that the abstract objects so defined are actually familiar objects?

19.8. Brouwer, the leader of the intuitionist school of mathematicians, is also known for major theorems in topology, including the invariance of geometric dimension under homeomorphisms and the *Brouwer fixed-point theorem*, which asserts that for any continuous mapping f of a closed disk into itself there is a point x such that $x = f(x)$. To prove this theorem, suppose there is a continuous mapping f for which $f(x) \neq x$ at every point x . Construct a continuous mapping g by drawing a line from $f(x)$ to x and extending it to the point $g(x)$ at which it meets the boundary circle (see Fig. 1). Then $g(x)$ maps the disk continuously onto its boundary circle and leaves each point of the boundary circle fixed. Such a continuous mapping is intuitively impossible (imagine stretching the entire head of a drum onto the rim without moving any point already on the rim and without tearing the head) and can be shown rigorously to be impossible (the disk and the circle have different homotopy groups). How can you explain the fact that the champion of intuitionism produced theorems that are not intuitionistically valid?

19.9. A naive use of the formula for the sum of the geometric series $1/(1+x) = 1 - x + x^2 - x^3 + \cdots$ seems to imply that $1 - 5 + 25 - 125 + \cdots = 1/(1+5) = 1/6$. Nineteenth-century analysts rejected this use of infinite series and confined themselves to series that converge in the ordinary sense. However, Kurt Hensel (1861–1941) showed in 1905 that it is possible to define a notion of distance (the p -adic metric) by saying that an integer is close to zero if it is divisible by a large power of the prime number p (in the present case, $p = 5$). Specifically, the distance from m to 0 is given by $d(m, 0) = 5^{-k}$, where 5^k divides m but 5^{k+1} does not divide m . The distance between 0 and the rational number $r = m/n$ is then by definition $d(m, 0)/d(n, 0)$. Show that $d(1, 0) = 1$. If the distance between two rational numbers r and s is defined to be $d(r - s, 0)$, then in fact the series just mentioned does converge to $\frac{1}{6}$ in the sense that $d(S_n, \frac{1}{6}) \rightarrow 0$, where S_n is the n th partial sum.

What does this historical experience tell you about the truth or falsity of mathematical statements? Is there an “understood context” for every mathematical statement that can never be fully exhibited, so that certain assertions will be *verbally* true in some contexts and verbally false in others, depending on the meaning attached to the terms?

19.10. Are there true but unknowable propositions in everyday life? Suppose that your class meets on Monday, Wednesday, and Friday. Suppose also that your instructor announces one Friday afternoon that you will be given a surprise exam at

one of the regular class meetings the following week. One of the brighter students then reasons as follows. The exam will not be given on Friday, since if it were, having been told that it would be one of the three days, and not having had it on Monday or Wednesday, we would know on Thursday that it was to be given on Friday, and so it wouldn't be a surprise. Therefore it will be given on Monday or Wednesday. But then, since we *know* that it can't be given on Friday, it also can't be given on Wednesday. For if it were, we would know on Tuesday that it was to be given on Wednesday, and again it wouldn't be a surprise. Therefore it must be given on Monday, we know that now, and therefore it isn't a surprise. Hence it is impossible to give a surprise examination next week.

Obviously something is wrong with the student's reasoning, since the instructor can certainly give a surprise exam. Most students, when trying to explain what is wrong with the reasoning, are willing to accept the first step. That is, they grant that it is impossible to give a *surprise* exam on the *last* day of an assigned window of days. Yet they balk at drawing the conclusion that this argument implies that the originally next-to-last day must thereby become the last day. Notice that, if the professor had said nothing to the students, it would be possible to give a surprise exam on the last day of the window, since the students would have no way of knowing that there was any such window. The conclusion that the exam cannot be given on Friday therefore does not follow from assuming a surprise exam within a limited window alone, but rather from these assumptions supplemented by the following proposition: *The students know that the exam is to be a surprise and they know the window in which it is to be given.*

This fact is apparent if you examine the student's reasoning, which is full of statements about what the students *would know*. Can they truly *know* a statement (even a true statement) if it leads them to a contradiction?

Explain the paradox in your own words, deciding whether the exam would be a surprise if given on Friday. Can the paradox be avoided by saying that the conditions under which the exam is promised are true but the students cannot *know* that they are true?

How does this puzzle relate to Gödel's incompleteness result?

Literature

Note: CSHPM/SCHPM = Canadian Society for the History and Philosophy of Mathematics/Société Canadienne d'Histoire et Philosophie des Mathématiques

- Adler, Ada, 1971. *Suidae Lexicon*, Teubner-Verlag, Stuttgart.
- Ağargün, Ahmet G.; Fletcher, Colin R., 1994. "Al-Farisi and the fundamental theorem of arithmetic," *Historia Mathematica*, **21**, No. 2, 162–173.
- Allman, George Johnston, 1889. *Greek Geometry from Thales to Euclid*, Longmans, Green & Co., London.
- Amir-Moez, Ali R., 1959. "Discussion of difficulties in Euclid by Omar ibn Abraham al-Khayyami," *Scripta Mathematica*, **XXIV**, 275–303.
- Amir-Moez, Ali R., 1963. "A paper of Omar Khayyam," *Scripta Mathematica*, **XXVI**, No. 4, 323–337.
- Andrews, George E., 1979. "An introduction to Ramanujan's 'lost' notebook," *American Mathematical Monthly*, **86**, No. 2, 89–108.
- Ang Tian-Se; Swetz, Frank J., 1986. "A Chinese mathematical classic of the third century: *The Sea Island Mathematical Manual* of Liu Hui," *Historia Mathematica*, **13**, No. 2, 99–117.
- Aschbacher, Michael, 1981. "The classification of the finite simple groups," *The Mathematical Intelligencer*, **3**, No. 2, 59–65.
- Ascher, Marcia, 1991. *Ethnomathematics*, Brooks/Cole, New York.
- Ascher, Marcia, 1992. "Before the conquest," *Mathematics Magazine*, **65**, No. 4, 211–218.
- Ascher, Marcia, 1995. "Models and maps from the Marshall Islands: a case in ethnomathematics," *Historia Mathematica*, **22**, 347–370.
- Ascher, Marcia, 1997. "Malagasy *Sikidy*: a case in ethnomathematics," *Historia Mathematica*, **24**, No. 4, 376–395.
- Ascher, Marcia; Ascher, Robert, 1997. *Mathematics of the Incas: Code of the Quipu*, Dover, Mineola, NY.
- Ayoub, R. 1980. "Paolo Ruffini's contributions to the quintic," *Archive for History of Exact Sciences*, **23**, No. 3, 253–277.
- Baatz, Simon, 1991. "'Squinting at Silliman': scientific periodicals in the early American Republic, 1810–1833," *Isis*, **82**, 223–244.
- Bag, Amulya Kumar, 1966. "Binomial theorem in ancient India," *Indian Journal of History of Science*, **1**, No. 1, 68–74.
- Bagheri, Mohammad, 1997. "A newly found letter of al-Kashi on scientific life in Samarkand," *Historia Mathematica*, **24**, No. 3, 241–256.

- Baigozhina, G. O., 1995. "On the classification principle of the problems in Abu-Kamil's *Book of Indeterminate Problems*," *Istoriko-Matematicheskie Issledovaniya*, **1** (36), 61–66 (Russian).
- Baker, J. N. L., 1948. "Mary Somerville and geography in England," *The Geographical Journal*, **111**, No. 4/6, 207–222.
- Baltzer, R., ed., 1885. *August Ferdinand Möbius, Gesammelte Werke*, S. Hirzel, Leipzig.
- Bashmakova, I. G.; Smirnova, G. S., 1997. "The origin and development of algebra," in: B. V. Gnedenko, ed., *Essays on the History of Mathematics*, Moscow University Press, pp. 94–246 (Russian). English translation published separately as *The Beginnings and Evolution of Algebra*, A. Shenitzer (transl.), Mathematical Association of America, Oberlin, OH, 2000.
- Beckers, Danny J., 1999. "Lagrange in the Netherlands: Dutch attempts to obtain rigor in calculus, 1797–1840," *Historia Mathematica*, **26**, No. 3, 234–238.
- Bedini, Silvio A., 1972. *The Life of Benjamin Banneker*, Charles Scribner's Sons, New York.
- Beman, Wooster Woodruff; Smith, David Eugene, 1930. *Famous Problems of Elementary Geometry*, G. E. Stechert & Co., New York.
- Berggren, J. L., 1989. "Abu Sahl al-Kuhi: what the manuscripts say," *Proceedings of the 15th annual meeting of the CSHPM/SCHPM*, Université Laval, Montréal, Québec, pp. 31–48.
- Berggren, J. L., 1990. "Greek and Islamic elements in Arabic mathematics," *Proceedings of the 16th annual Meeting of the CSHPM/SCHPM*, University of Victoria, Victoria, British Columbia, pp. 25–38.
- Berggren, J. L., 2002. "The transmission of Greek geometry to medieval Islam," *CUBO*, **4**, No. 2, 1–13.
- Bernal, Martin, 1992. "Animadversions on the origins of western science," *Isis*, **83**, No. 4, 596–607.
- Bernays, Paul, ed., 1971. *Foundations of Geometry*, by David Hilbert, translated by Leo Unger, Open Court, La Salle, IL.
- Berndt, Bruce C., ed., 1985. *Ramanujan's Notebooks*, 5 vols., Springer-Verlag, New York.
- Betti, E. 1852. "Sulla risoluzione delle equazioni algebriche," *Tortolini Annali*, **III**, 49–51.
- Biggs, N. L., 1979. "The roots of combinatorics," *Historia Mathematica*, **6**, No. 2, 109–136.
- Biggs, N. L., 1981. "T. P. Kirkman, mathematician," *Bulletin of the London Mathematical Society*, **13**, 97–120.
- Billard, Lynne, 1991. "The past, present, and future of academic women in the mathematical sciences," *Notices of the American Mathematical Society*, **38**, No. 7, 707–714.
- Blackwell, Richard, transl., 1986. *Christiaan Huygens' The Pendulum Clock*, Iowa State University Press, Ames, IA.

- Blum, Lenore, 1991. "A brief history of the Association for Women in Mathematics: the presidents' perspectives," *Notices of the American Mathematical Society*, **38**, No. 7, 738–754.
- Bochenski, I. M., 1961. *A History of Formal Logic*, University of Notre Dame Press.
- Bochner, Salomon, 1974. "Mathematical reflections," *American Mathematical Monthly*, **81**, No. 8, 827–840.
- Boncompagni, Baldassare, 1854. *Intorno ad alcune opere matematica notizie di Leonardo, matematico del secolo decimoterzo*, Tipografia Delle Belle Arti, Rome.
- Bottazzini, Umberto, 1986. *The Higher Calculus: A History of Real and Complex Analysis from Euler to Weierstrass*, Springer-Verlag, New York.
- Boyer, Carl B., 1949. *The History of the Calculus and its Conceptual Development*, Hafner, New York. Reprint: Dover, New York, 1959.
- Brentjes, Sonja; Hogendijk, Jan P., 1989. "Notes on Thabit ibn Qurra and his rule for amicable numbers," *Historia Mathematica*, **16**, No. 4, 373–378.
- Bretschneider, Carl Anton, 1870. *Die Geometrie und die Geometer vor Euklides. Ein historischer Versuch*, Teubner, Leipzig. Reprint: M. Sändig, Wiesbaden, 1968.
- Brown, James Robert, 1999. *Philosophy of Mathematics: An Introduction to the World of Proofs and Pictures*, Routledge, New York.
- Buck, R. C., 1980. "Sherlock Holmes in Babylon," *The American Mathematical Monthly*, **87**, No. 5, 335–345.
- Burington, Richard Stevens, 1958. *Handbook of Mathematical Tables and Formulas*, Handbook Publishers, Sandusky, OH.
- Butzmann, Hans, 1970. *Codex Agrimensorum Romanorum: Codex Arcerianusa der Herzog-August-Bibliothek zu Wolfenbüttel*, A. W. Sijthoff, Lugduni Batavorum.
- Bychkov, S. N., 2001. "Egyptian geometry and Greek science," *Istoriko-Matematicheskie Issledovaniya*, **6** (41), 277–284 (Russian).
- Cantor, Moritz, 1880. *Vorlesungen über Geschichte der Mathematik*, Vol. 1, Teubner, Leipzig.
- Cauchy, A.-L., 1815. "Mémoire sur le nombre des valeurs qu'une fonction peut acquérir," *Journal de l'École Polytechnique*, *XVII^e Cahier*, Tome X = *Œuvres Complètes*, Tome 13, pp. 64–90. Gauthier-Villars, Paris.
- Caveing, Maurice, 1985. "La tablette babylonienne AO 17264 du Musée du Louvre et le problème des six frères," *Historia Mathematica*, **12**, No. 1, 6–24.
- Chace, A. B., Bull, L., Manning, H. P., and Archibald, R. C., 1927. *The Rhind Mathematical Papyrus*, Mathematical Association of America, Oberlin, OH, Vol. 1.
- Chaikovskii, Yu. V., 2001. "What is probability? The evolution of the concept (from antiquity to Poisson)," *Istoriko-Matematicheskie Issledovaniya*, **6** (41), 34–56 (Russian).
- Chaucer, Geoffrey, 1391. *A Treatise on the Astrolabe*, in: F. N. Robinson, ed., *The Works of Geoffrey Chaucer*, 2nd ed., Houghton Mifflin, Boston, 1957, pp. 545–563.
- Chemla, Karine, 1991. "Theoretical aspects of the Chinese algorithmic tradition (first to third century)," *Historia Scientiarum*, No. 42, 75–98.
- Christianidis, Jean, 1998. "Une interprétation byzantine de Diophante," *Historia Mathematica*, **25**, No. 1, 22–28.

- Clagett, Marshall, 1968. *Nicole Oresme and the Medieval Geometry of Qualities and Motions*, University of Wisconsin Press, Madison, WI.
- Clark, Walter Eugene, ed., 1930. *The Aryabhatiya of Aryabhata*, University of Chicago Press, Chicago.
- Clayton, Peter A., 1994. *Chronicle of the Pharaohs*, Thames & Hudson, London.
- Closs, Michael P., ed., 1986. *Native American Mathematics*, University of Texas Press, Austin, TX.
- Closs, Michael P., 1992. "Ancient Maya mathematics and mathematicians," *Proceedings of the 18th Annual Meeting of the CSHPM/SCHPM*, University of Prince Edward Island, Charlottetown, P.E.I., pp. 1–13.
- Coe, Michael D., 1973. *The Maya Scribe and His World*, The Grolier Club, New York.
- Colebrooke, Henry Thomas, 1817. *Algebra with Arithmetic and Mensuration from the Sanscrit of Brahmagupta and Bhascara*, J. Murray, London.
- Colson, F. H., 1926. *The Week: An Essay on the Origin and Development of the Seven-Day Cycle*, Greenwood Press, Westport, CT.
- Coolidge, Julian Lowell, 1940. *A History of Geometrical Methods*, Clarendon Press, Oxford.
- Craik, Alex D. D., 1999. "Calculus and analysis in early nineteenth-century Britain: the work of William Wallace," *Historia Mathematica*, **26**, No. 3, 239–267.
- Crossley, John N.; Henry, Alan S., 1990. "Thus spake al-Khwārizmī: A translation of the text of Cambridge University Library Ms. ii.vi.5," *Historia Mathematica*, **17**, No. 2, 103–131.
- Cullen, Christopher, 1996. *Astronomy and Mathematics in Ancient China: The Zhou Bi Suan Jing*, Cambridge University Press.
- al-Daffa, Ali Abdullah, 1977. *The Muslim Contribution to Mathematics*, Humanities Press, Atlantic Highlands, NJ.
- Dahan, Amy, 1980. "Les travaux de Cauchy sur les substitutions; Étude de son approche du concept de groupe," *Archive for History of Exact Sciences*, **23**, No. 4, 279–319.
- Dahan-Dalmédico, Amy, 1987. "Mécanique et théorie des surfaces: les travaux de Sophie Germain," *Historia Mathematica*, **14**, No. 4, 347–365.
- Dauben, Joseph W., 1996. "Mathematics at the University of Toronto: Abraham Robinson in Canada (1951–1957)," in: Dauben, Folkerts, Knobloch, and Wussing, *History of Mathematics: States of the Art*, Academic Press, New York.
- Dauben, Joseph W.; Scriba, Christoph J., eds., 2002. *Writing the History of Mathematics: Its Historical Development*, Birkhäuser, Boston.
- David, H. A.; Edwards, A. W. F., 2001. *Annotated Readings in the History of Statistics*, Springer-Verlag, New York.
- Davis, Margaret Daly, 1977. *Piero Della Francesca's Mathematical Treatises: the Tratto d'abaco and Libellus de quinque corporibus regularibus*, Longe Editore, Ravenna.
- Davis, Philip J.; Hersh Reuben, 1986. *Descartes' Dream: The World According to Mathematics*, Harcourt Brace Jovanovich, New York.

- Deakin, Michael, 1994. "Hypatia and her mathematics," *American Mathematical Monthly*, **101**, No. 3, 234–243.
- Dean, Nathaniel, ed., 1996. *African Americans in Mathematics. DIMACS Workshop, June 26–28, 1996*, American Mathematical Society, Providence, RI.
- Detlefsen, Michael, 2001. "What does Gödel's second theorem say?" *Philosophia Mathematica* (3), **9**, No. 1, 37–71.
- Detlefsen, Michael; Erlandson, Douglas K.; Heston, J. Clark; Young, Charles M., 1975. "Computation with Roman numerals," *Archive for History of Exact Science*, **15**, No. 2, 141–148.
- Dick, Auguste, 1981. *Emmy Noether, 1882–1935*, translated by H.I. Blocher. Birkhäuser, Boston.
- Dickson, Leonard Eugene, 1919. *History of the Theory of Numbers I: Divisibility and Primality*, Carnegie Institute, Washington, DC. Reprint: Chelsea, New York, 1966.
- Dickson, Leonard Eugene, 1920. *History of the Theory of Numbers II: Diophantine Analysis*, Carnegie Institute, Washington, DC. Reprint: Chelsea, New York, 1966.
- Dickson, Leonard Eugene, 1923. *History of the Theory of Numbers III: Quadratic and Higher Forms*, Carnegie Institute, Washington, DC. Reprint: Chelsea, New York, 1966.
- Diels, Hermann, 1951. *Die Fragmente der Vorsokratiker*, 6th corrected edition, (Walther Kranz, ed.), Weidmann, Berlin.
- Dijksterhuis, E. J., 1956. *Archimedes*, Munksgård, Copenhagen.
- Dilke, O. A. W., 1985. *Greek and Roman Maps*, Cornell University Press, Ithaca, NY.
- D'ooge, Martin Luther, 1926, translator. *Introduction to Arithmetic* (Nicomachus of Gerasa), Macmillan, New York.
- Dorofeeva, A. V., 1998. "The calculus of variations," in: *Mathematics of the 19th Century*, Birkhäuser, Basel, pp. 197–260.
- Duren, Peter, 1989. *A Century of Mathematics in America* (3 Vols.), American Mathematical Society, Providence, RI.
- Dutka, Jacques, 1988. "On the Gregorian revision of the Julian calendar," *Mathematical Intelligencer*, **10**, No. 1, 56–64.
- Dzielska, Maria, 1995. *Hypatia of Alexandria*, Harvard University Press, Cambridge, MA.
- Edwards, H. M., 1974. *Riemann's Zeta Function*, Academic Press, New York.
- Edwards, H. M., 1977. *Fermat's Last Theorem*, Springer-Verlag, New York.
- Ehrman, Esther, 1986. *Mme du Châtelet*, Berg Publishers, Leamington Spa, Warwickshire.
- Engel, Friedrich; Heegaard, Poul, eds., 1960. *Sophus Lie, Gesammelte Abhandlungen*, Teubner, Leipzig.
- Erdős, P.; Dudley, U., 1983. "Some remarks and problems in number theory related to the work of Euler," *Mathematics Magazine*, **56**, No. 5, 292–298.
- Euler, L., 1732. "De formes radicum aequationum cuiusque ordinis coniectatio," *Commentarii Academiae Petropolitanae*, **6** (1738), p. 216.

- Euler, L., 1744. *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes*, Bousquet & Co., Lausanne.
- Euler, L., 1749. "Recherches sur les racines imaginaires des équations," *Histoire de l'Académie des Sciences de Berlin*, p. 222.
- Euler, L., 1762. "De resolutione aequationum cuiusque gradus," *Novi Commentarii Academiae Petropolitanae* 9, p. 70.
- Farrar, John, 1826. *Elements of Electricity, Magnetism, and Electro-magnetism*, Hilliard and Metcalf, Cambridge, MA.
- Feigenbaum, L., 1994. "Infinite series and solutions of ordinary differential equations, 1670–1770," in: *Companion Encyclopedia of the History and Philosophy of the Mathematical Sciences*, Vol. 1, Routledge, London and New York, pp. 504–519.
- Feingold, Mordechai, 1993. "Newton, Leibniz, and Barrow too: an attempt at a reinterpretation," *Isis*, 84, No. 2, 310–338.
- Feit, Walter, 1977. "A mathematical visit to China, May 1976," *Notices of the American Mathematical Society*, 24, No. 2, 110–113.
- Fennema, Elizabeth; Leder, Gilah C., eds., 1990. *Mathematics and Gender*, Teachers College Press, New York.
- Ferreirós, José, 1995. "'What fermented in me for years': Cantor's discovery of transfinite numbers," *Historia Mathematica*, 22, 33–42.
- Field, J. V.; Gray, J. J., 1987. *The Geometrical Work of Girard Desargues*, Springer-Verlag, Berlin.
- Fitzgerald, Augustine, transl., 1926. *The Letters of Synesius of Cyrene*, Oxford University Press.
- Fletcher, Colin R., 1989. "Fermat's theorem," *Historia Mathematica*, 16, No. 2, 149–153.
- Folkerts, Menso, 1970. *Ein neuer Text des Euclides latinus*, H. A. Gerstenberg, Hildesheim.
- Folkerts, Menso, 1971. *Anonyme lateinische Euklidbearbeitungen aus dem 12. Jahrhundert*, Österreichische Akademie der Wissenschaften, Mathematisch-Naturwissenschaftliche Klasse, Denkschriften, 116, erste Abhandlung.
- Fowler, David, 1992. "Dedekind's theorem: $\sqrt{2} \times \sqrt{3} = \sqrt{6}$," *The American Mathematical Monthly*, 99, No. 8, 725–733.
- Fowler, David, 1998. *The Mathematics of Plato's Academy*, 2nd ed., Clarendon Press, Oxford.
- Franchella, Miriam, 1995. "L. E. J. Brouwer: toward intuitionistic logic," *Historia Mathematica*, 22, No. 3, 304–322.
- Franci, Rafaella, 1988. "Antonio de' Mazzinghi, an algebraist of the fourteenth century," *Historia Mathematica*, 15, No. 3, 240–249.
- Fraser, Craig, 1987. "Joseph-Louis Lagrange's algebraic vision of the calculus," *Historia Mathematica*, 14, No. 1, 38–53.
- Fraser, Craig, 1993. "A history of Jacobi's theorem in the calculus of variations," *Proceedings of the 19th Annual Meeting of the CSPHM/SCHPM*, Carleton University, Ottawa, Ontario, pp. 168–185.
- Freidel, David; Schele, Linda; Parker, Joy, 1993. *Maya Cosmos: Three Thousand Years on the Shaman's Path*, William Morrow, New York.

- Friberg, Jöran, 1981. "Methods and traditions of Babylonian mathematics: Plimpton 322, Pythagorean triples and the Babylonian triangle parameter equations," *Historia Mathematica*, 8, No. 3, 277–318.
- von Fritz, Kurt, 1945. "The discovery of incommensurability by Hippasus of Metapontum," *Annals of Mathematics*, 46, 242–264.
- Fu Daiwie, 1991. "Why did Lui Hui fail to derive the volume of a sphere?," *Historia Mathematica*, 18, No. 3, 212–238.
- Fukagawa Hidetoshi; Pedoe, D., 1989. *Japanese Temple Geometry Problems: San Gaku*, Winnipeg, Manitoba.
- Fuson, Karen, 1988. *Children's Counting and Concepts of Number*, Springer-Verlag, New York.
- Garciadiego, Alejandro, 2002. "Mexico," in: Dauben and Scriba, (2002, pp. 256–263).
- Gauss, Carl Friedrich, 1799. "Demonstratio nova theorematis omnem functionem algebraicam rationalem integram unius variabilis in factores reales primi vel secundi gradus resolvi posse," in: *Werke*, Vol. 3, Königlichen Gesellschaft der Wissenschaften, Göttingen, 1866, pp. 3–31.
- Gauss, Carl Friedrich, 1965. *General Investigations of Curved Surfaces*, translated from the Latin and German by Adam Hiltebeitel and James Morehead, Raven Press, Hewlett, NY.
- Geijsbeek, John B., ed. and transl., 1914. *Ancient Double-Entry Bookkeeping. Lucas Pacioli's Treatise (A.D. 1494—The Earliest Known Writer on Bookkeeping)*, Denver, CO.
- Gerdes, Paulus, 1985. "Three alternate methods of obtaining the ancient Egyptian formula for the area of a circle," *Historia Mathematica*, 12, No. 3, 261–268.
- Gerhardt, C. I., ed., 1971. *Leibniz: Mathematische Schriften*, G. Olms Verlag, Hildesheim.
- Gericke, Helmut, 1996. "Zur Geschichte der negativen Zahlen," in: Dauben, Folkerts, Knobloch, and Wussing, eds., *History of Mathematics: States of the Art*, Academic Press, New York, pp. 279–306.
- Gericke, Helmut; Vogel, Kurt, 1965. *De Thiende von Simon Stevin*, Akademische Verlagsgesellschaft, Frankfurt am Main.
- Gillings, Richard J., 1972. *Mathematics in the Time of the Pharaohs*, MIT Press, Cambridge, MA. Reprint: Dover, New York, 1982.
- Gold, David; Pingree, David, 1991. "A hitherto unknown Sanskrit work concerning Mādhava's derivation of the power series for sine and cosine," *Historia Scientiarum*, No. 42, 49–65.
- Goldstine, Herman H., 1980. *A History of the Calculus of Variations from the 17th Through the 19th Century*, Springer-Verlag, New York.
- Gottwald, Siegfried; Ilgauds, Hans-Joachim; Schlote, Karl-Heinz, eds., 1990. *Lezikon Bedeutender Mathematiker*, Bibliographisches Institut, Leipzig.
- Gould, James L.; Gould, Carol Grant, 1995. *The Honey Bee*, Scientific American Library, New York.
- Gow, James, 1884. *A Short History of Greek Mathematics*. Reprint: Chelsea, New York, 1968.

- Grabner, Judith, 1995. "Descartes and problem-solving," *Mathematics Magazine*, **68**, No. 2, 83–97.
- Grattan-Guinness, Ivor, 1972. "A mathematical union: William Henry and Grace Chisholm Young," *Annals of Science*, **29**, No. 2, 105–186.
- Grattan-Guinness, Ivor, 1975. "Mathematical bibliography for W.H. and G.C. Young," *Historia Mathematica*, **2**, 43–58.
- Grattan-Guinness, Ivor, 1990. *Convolutions in French Mathematics, 1800–1840*, Birkhäuser, Basel.
- Grattan-Guinness, Ivor, 2000. *The Search for Mathematical Roots, 1870–1940: Logics, Set Theories, and the Foundations of Mathematics from Cantor through Russell to Gödel*, Princeton University Press, Princeton, NJ.
- Grattan-Guinness, Ivor, 2004. "History or heritage? An important distinction in mathematics and for mathematics education," *American Mathematical Monthly*, **111**, 1–12.
- Gray, J. J., 1989. *Ideas of Space: Euclidean, Non-Euclidean, and Relativistic*, 2nd ed., Clarendon Press, Oxford.
- Gray, Robert, 1994. "Georg Cantor and transcendental numbers," *The American Mathematical Monthly*, **101**, No. 9, 819–832.
- Green, Judy; LaDuke, Jeanne, 1987. "Women in the American mathematical community: the pre-1940 Ph.D.'s," *The Mathematical Intelligencer*, **9**, No. 1, 11–23.
- Greenleaf, Benjamin, 1876. *New Practical Arithmetic; in Which the Science and Its Applications Are Simplified by Induction and Analysis*, Leach, Shewell, and Sanborn, Boston and New York.
- Grinstein, Louise S.; Campbell, Paul J., 1982. "Anna Johnson Pell Wheeler: Her life and work," *Historia Mathematica*, **9**, No. 1, 37–53.
- Grosholz, Emily, 1987. "Two Leibnizian manuscripts of 1690 concerning differential equations," *Historia Mathematica*, **14**, No. 1, 1–37.
- Guizal, Brahim; Dudley, John, 2002. "Ibn Sahl: Inventeur de la loi de la réfraction," *Revue pour la science*, No. 301, November 2002.
- Guo Shuchung, 1992. "Guo Shuchung's edition of the *Jiu Zhang Suan Shu*," *Historia Mathematica*, **19**, No. 2, 200–202.
- Gupta, R. C., 1989. "Sino-Indian interaction and the great Chinese Buddhist astronomer-mathematician I-Hsing (A.D. 683–727)," *Bulletin of the Indian Society for History of Mathematics*, **11**, Nos. 1–4, 38–49.
- Gupta, R. C., 1991. "On the volume of a sphere in ancient India," *Historia Scientiarum*, No. 42, 33–44.
- Gupta, R. C., 1994a. "Six types of Vedic mathematics," *Bulletin of the Indian Society for History of Mathematics*, **16**, Nos. 1–4, 5–15.
- Gupta, R. C., 1994b. "A circulatory rule from the *Agni Purāṇa*," *Bulletin of the Indian Society for History of Mathematics*, **16**, Nos. 1–4, 53–56.
- Gustafson, W. H.; Halmos, P. R.; Moolgavkar, S. H.; Wheeler, W. H.; Ziemer, W. P., 1976. "American mathematics from 1940 to the day before yesterday," *The American Mathematical Monthly*, **83**, No. 7, 503–516.

- Hadamard, J., 1896. "Sur la distribution des zéros de la fonction $\zeta(s)$ et ses conséquences arithmétiques," *Bulletin de la Société Mathématique de France*, **24**, 199–220.
- Hairetdinova, N. G., 1986. "On spherical trigonometry in the medieval Near East and in Europe," *Historia Mathematica*, **13**, No. 2, 136–146.
- Hamilton, W. R., 1837. "On the argument of Abel, respecting the impossibility of expressing a root of any general equation above the fourth degree, by any finite combination of radicals and rational functions," *Transactions of the Royal Irish Academy*, **XVIII** (1839), 171–259; H. Halberstam and R. Ingram, eds., *The Mathematical Papers of Sir William Rowan Hamilton*, Vol. III, pp. 517–569.
- Hari, K. Chandra, 2002. "Genesis and antecedents of Āryabhaṭīya," *Indian Journal of History of Science*, **37**, No. 2, 101–113.
- Harrison, Jenny, 1991. "The Escher staircase," *Notices of the American Mathematical Society*, **38**, No. 7, 730–734.
- al-Hassan, Ahmad Y.; Hill, Donald R., 1986. *Islamic Technology: An Illustrated History*, Cambridge University Press.
- Hawkins, Thomas, 1989. "Line geometry, differential equations, and the birth of Lie's theory of groups," in: David Rowe and John McCleary, eds, *The History of Modern Mathematics*, Vol. 1, Academic Press, New York.
- Hayashi Takao, 1987. "Varahamihira's pandiagonal magic square of the order four," *Historia Mathematica*, **14**, No. 2, 159–166.
- Hayashi Takao, 1991. "A note on Bhāskara I's rational approximation to sine," *Historia Scientiarum*, No. 42, pp. 45–48.
- Heath, T. L., 1910. *Diophantus of Alexandria: A Study in the History of Greek Algebra*, 2nd ed., Cambridge University Press, 1910.
- Heath, T. L., 1897–1912. *The Works of Archimedes Edited in Modern Notation with Introductory Chapters*, Reprint: Dover, New York, 1953.
- Heath, T. L., 1921. *A History of Greek Mathematics*, Clarendon Press, Oxford.
- Henrion, Claudia, 1991. "Merging and emerging lives: women in mathematics," *Notices of the American Mathematical Society*, **38**, No. 7, 724–729.
- Henrion, Claudia, 1997. *Women in Mathematics: The Addition of Difference*, Indiana University Press, Bloomington, IN.
- Heppenheimer, T. A., 1990. "How von Neumann showed the way," *American Heritage of Invention and Technology*, **6**, No. 2, 8–16.
- Hersh, Reuben, 1997. *What Is Mathematics, Really?* Oxford University Press, New York.
- Heyde, C. C.; Seneta, E., 1977. *I. J. Bienaymé: Statistical Theory Anticipated*, Springer-Verlag, New York.
- Hilbert, David, 1971. *Foundations of Geometry*, translated by Leo Unger, Open Court, LaSalle, IL.
- Hogendijk, Jan P., 1985. "Thabit ibn Qurra and the pair of amicable numbers 17296, 18416," *Historia Mathematica*, **12**, No. 3, 269–273.
- Hogendijk, Jan P., 1989. "Sharaf al-Din al-Tusi on the number of positive roots of cubic equations," *Historia Mathematica*, **16**, No. 1, 69–85.

- Hogendijk, Jan P., 1991. "Al-Khwarizmi's table of the 'sine of the hours' and underlying sine table," *Historia Scientiarum*, No. 42, pp. 1–12.
- Hogendijk, Jan P., 2002. "The surface area of the bicylinder and Archimedes' Method," *Historia Mathematica*, **29**, No. 1, 199–203.
- Homann, Frederick A., 1987. "David Rittenhouse: logarithms and leisure," *Mathematics Magazine*, **60**, No. 1, 15–20.
- Homann, Frederick A., 1991. *Practical Geometry: Practica Geometriae, attributed to Hugh of St. Victor*, Marquette University Press, Milwaukee, WI.
- Høyrup, Jens, 2002. "A note on Old Babylonian computational techniques," *Historia Mathematica*, **29**, No. 2, 193–198.
- Hughes, Barnabas, 1981. *De numeris datis* (Jordan de Nemore), University of California Press, Berkeley, CA.
- Hughes, Barnabas, 1989. "The arithmetical triangle of Jordanus de Nemore," *Historia Mathematica*, **16**, No. 3, 213–223.
- Hultsch, F., ed., 1965. *Pappi Alexandrini Collectionis*, Vol. 1, Verlag Adolf M. Hakkert, Amsterdam.
- Ifrah, Georges, 2000. *The Universal History of Numbers: From Prehistory to the Invention of the Computer*, translated from the French by David Bellos, Wiley, New York.
- Il'ina, E. A., 2002. "On Euclid's *Data*," *Istoriko-Matematicheskie Issledovaniya*, **7** (42), 201–208 (Russian).
- Indorato, Luigi; Nastasi, Pietro, 1989. "The 1740 resolution of the Fermat–Descartes controversy," *Historia Mathematica*, **16**, No. 2, 137–148.
- Ivić, Aleksandar, 1985. *The Riemann Zeta-Function: Theory and Applications*, Dover, New York.
- Ivins, W. M., 1947. "A note on Desargues' theorem," *Scripta Mathematica*, **13**, 203–210.
- Jackson, Allyn, 1991. "Top producers of women mathematics doctorates," *Notices of the American Mathematical Society*, **18**, No. 7, 715–720.
- Jacobs, Konrad; Utz, Heinrich, 1984. "Erlangen programs," *The Mathematical Intelligencer*, **6**, No. 1.
- James, Portia P., 1989. *The Real McCoy: African-American Invention and Innovation, 1619–1930*, Smithsonian Institution Press, Washington, DC.
- Jami, Catherine, 1988. "Western influence and Chinese tradition in an eighteenth century Chinese mathematical work," *Historia Mathematica*, **15**, No. 4, 311–331.
- Jami, Catherine, 1991. "Scholars and mathematical knowledge during the late Ming and early Qing," *Historia Scientiarum*, No. 42, 95–110.
- Jentsch, Werner, 1986. "Auszüge aus einer unveröffentlichten Korrespondenz von Emmy Noether und Hermann Weyl mit Heinrich Brandt," *Historia Mathematica*, **13**, No. 1, 5–12.
- Jha, V. N., 1994. "Indeterminate analysis in the context of the Mahāsiddhānta of Āryabhaṭa II," *Indian Journal of History of Science*, **29**, No. 4, 565–578.
- Jones, Alexander, 1991. "The adaptation of Babylonian methods in Greek numerical astronomy," *Isis*, **82**, No. 313, 441–453.

- Kasir, Daoud, 1931. *The Algebra of Omar Khayyam*, Teachers College, Columbia University Contributions to Education, No. 385, New York.
- Kawahara Hideki, 1991. "World-View of the *Santong-Li*," *Historia Scientiarum*, No. 42, 67–73.
- Kazdan, Jerry L., 1986. "A visit to China," *The Mathematical Intelligencer*, 8, No. 4, 22–32.
- Keith, Natasha; Keith, Sandra Z., 2000. Review of Claudia Henrion's *Women in Mathematics*, *Humanistic Mathematics Network Journal*, Issue 22, 26–30.
- Kenschaft, Patricia, 1981. "Black women in mathematics in the United States," *The American Mathematical Monthly*, 88, No. 8, 592–604.
- Kimberling, C. H., 1972a. "Emmy Noether," *The American Mathematical Monthly*, 79, No. 2, 136–149.
- Kimberling, C. H., 1972b. "Addendum to 'Emmy Noether'," *The American Mathematical Monthly*, 79, No. 7, 755.
- King, R. Bruce, 1996. *Beyond the Quartic Equation*, Birkhäuser, Boston.
- Kiro, S. N., 1967, "N. I. Lobachevskii and mathematics at Kazan' University," in: *History of Russian and Soviet Mathematics (Istoriya Otechestvennoi Matematiki)*, Vol. 2, Naukova Dumka, Kiev (Russian).
- Klein, Felix, 1884. *Lectures on the Icosahedron and the Solution of Equations of the Fifth Degree*, translated by George Gavin Morrice. Reprint: Dover, New York, 1956.
- Klein, Felix, 1926. *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*, Springer-Verlag, Berlin, 2 vols. Reprint: American Mathematical Society (Chelsea Publishing Company), Providence, RI, 1967.
- Klein, Jacob, 1933. *Plato's Trilogy: Theaetetus, the Sophist, and the Statesman*, University of Chicago Press, Chicago, 1977.
- Klein, Jacob, 1934–1936. *Greek Mathematical Thought and the Origin of Algebra*, translated by Eva Brann, MIT Press, Cambridge, MA, 1968.
- Klein, Jacob, 1965. *A Commentary on Plato's Meno*, University of North Carolina Press, Chapel Hill, NC.
- Kleiner, Israel, 1991. "Emmy Noether: Highlights of her life and work," *Proceedings of the 17th Annual Meeting of the CSHPM/SCHPM*, Queen's University, Kingston, Ontario, pp. 19–42.
- Knorr, Wilbur, 1975. *The Evolution of the Euclidean Elements*, Reidel, Boston.
- Knorr, Wilbur, 1976. "Problems in the interpretation of Greek number theory: Euclid and the 'fundamental theorem of arithmetic'," *Studies in the Historical and Philosophical Sciences*, 7, 353–368.
- Knorr, Wilbur, 1982. "Techniques of fractions in ancient Egypt," *Historia Mathematica*, 9, No. 2, 133–171.
- Koblitz, Ann Hibner, 1983. *A Convergence of Lives. Sophia Kovalevskaja: Scientist, Writer, Revolutionary*. Birkhäuser, Boston.
- Koblitz, Ann Hibner, 1984. "Sofia Kovalevskaja and the mathematical community," *The Mathematical Intelligencer*, 6, No. 1, 20–29.

- Koblitz, Ann Hibner, 1991. "Historical and cross-cultural perspectives on women in mathematics," *Proceedings of the 17th Annual Meeting of the CSHPM/SCHPM*, Queen's University, Kingston, Ontario, pp. 1-18.
- Koblitz, Neal, 1990. "Recollections of mathematics in a country under siege," *The Mathematical Intelligencer*, **12**, No. 3, 16-34.
- Koehler, Otto, 1937. *Bulletin of Animal Behavior*, No. 9. English translation in James R. Newman, ed., *The World of Mathematics*, vol. 1, Simon and Schuster, New York, 1956, pp. 491-492.
- Kowalewski, Gerhard, 1950. *Bestand und Wandel*, Oldenbourg, München.
- Kox, A. J.; Klein, Martin J.; Schulmann, Robert, eds., 1996. *The Collected Papers of Albert Einstein*, Princeton University Press, Princeton, NJ.
- Kracht, Manfred; Kreyszig, Erwin, 1990. "E. W. von Tschirnhaus: his role in early calculus and his work and impact on algebra," *Historia Mathematica*, **17**, No. 1, 16-35.
- Kreyszig, Erwin, 1993. "On the calculus of variations and its major influences on the mathematics of the first half of our century," *Proceedings of the 19th Annual Meeting of the CSHPM/SCHPM*, Carleton University, Ottawa, Ontario, pp. 119-149.
- Krupp, E. C., 1983. *Echoes of the Ancient Skies*, Harper & Row, New York.
- Kunoff, Sharon, 1990. "A curious counting/summation formula from the ancient Hindus," in: *Proceedings of the 16th Annual Meeting of the CSHPM/SCHPM*, University of Victoria, Victoria, British Columbia, pp. 101-107.
- Kunoff, Sharon, 1991. "Women in mathematics: Is history being rewritten?" *Proceedings of the 17th Annual Meeting of the CSHPM/SCHPM*, Queen's University, Kingston, Ontario, pp. 43-52.
- Kunoff, Sharon, 1992. "Some inheritance problems in ancient Hebrew literature," in: *Proceedings of the 18th Annual Meeting of the CSHPM/SCHPM*, University of Prince Edward Island, Charlottetown, P.E.I., pp. 14-20.
- Lagrange, J.-L., 1771. "Réflexions sur la résolution algébrique des équations," *Nouveaux mémoires de l'Académie Royale des Sciences et Belles-lettres de Berlin; Œuvres* (1869), Vol. 3, pp. 205-421.
- Lagrange, J.-L., 1795. *Lectures on Elementary Mathematics*, translated by Thomas J. McCormack, Open Court, Chicago, 1898.
- Lam Lay-Yong, 1994. "Jiu Zhang Suanshu (Nine Chapters on the Mathematical Art): An Overview," *Archive for History of Exact Sciences*, **47**, No. 1, 1-51.
- Lam Lay-Yong; Ang Tian-Se, 1986. "Circle measurements in ancient China," *Historia Mathematica*, **13**, No. 4, 325-340.
- Lam Lay-Yong; Ang Tian-Se, 1987. "The earliest negative numbers: how they emerged from a solution of simultaneous linear equations," *Archive for History of Exact Sciences*, **37**, 222-267.
- Lam Lay-Yong; Ang Tian-Se, 1992. *Fleeting Footsteps. Tracing the Conception of Arithmetic and Algebra in Ancient China*, World Scientific, River Edge, NJ.
- Lam Lay-Yong; Shen Kangsheng, 1985. "The Chinese concept of Cavalieri's principle and its applications," *Historia Mathematica*, **12**, No. 3, 219-228.

- Lasserre, François, 1964. *The Birth of Mathematics in the Age of Plato*, Hutchinson, London.
- Laugwitz, Detlef, 1987. "Infinitely small quantities in Cauchy's textbooks," *Historia Mathematica*, **14**, No. 3, 258–274.
- Laugwitz, Detlef, 1999. *Bernhard Riemann, 1826–1866: Turning Points in the Conception of Mathematics*, Birkhäuser, Boston.
- Levey, Martin, 1966. *The Algebra of Abū Kāmil*, University of Wisconsin Press, Madison, WI.
- Levey, Martin; Petruck, Marvin, 1965. *Kūshyār ibn Labbān: Principles of Hindu Reckoning*, University of Wisconsin Press, Madison, WI.
- Lewis, D. J., 1991. "Mathematics and women: the undergraduate school and the pipeline," *Notices of the American Mathematical Society*, **38**, No. 7, 721–723.
- Libbrecht, Ulrich, 1973. *Chinese Mathematics in the Thirteenth Century*, MIT Press, Cambridge, MA.
- Liebmann, Heinrich, 1904. *N. J. Lobatschewskijs imaginäre Geometrie und Anwendung der imaginären Geometrie auf einige Integrale*, Teubner, Leipzig.
- Li Yan; Du Shiran, 1987. *Chinese Mathematics: A Concise History*, translated by John N. Crossley and Anthony W.-C. Lun, Clarendon Press, Oxford.
- Mack, John, 1990. *Emil Torday and the Art of the Congo. 1900–1909*, University of Washington Press, Seattle, WA.
- Mackay, Alan L., 1991. *A Dictionary of Scientific Quotations*, Institute of Physics Publishing, Bristol and Philadelphia.
- Mallayya, V. Madhukar, 1997. "Arithmetic operation of division with special reference to Bhāskara II's *Līlāvātī* and its commentaries," *Indian Journal of History of Science*, **32**, No. 4, 315–324.
- Mancosu, Paolo, 1989. "The metaphysics of the calculus: a foundational debate in the Paris Academy of Sciences, 1700–1706," *Historia Mathematica*, **16**, No. 3, 224–248.
- Manheim, Jerome H., 1964. *The Genesis of Point Set Topology*, Macmillan, New York.
- Martzloff, Jean-Claude, 1982. "Li Shanlan (1811–1882) and Chinese traditional mathematics," *The Mathematical Intelligencer*, **14**, No. 4, 32–37.
- Martzloff, Jean-Claude, 1990. "A survey of Japanese publications on the history of Japanese traditional mathematics (*Wasan*) from the last 30 years," *Historia Mathematica*, **17**, No. 4, 366–373.
- Martzloff, Jean-Claude, 1993. "Éléments de réflexion sur les réactions chinoises à la géométrie euclidienne à la fin du XVIII^e siècle—Le *Jihe lunyue* {a} de Du Zhigeng {b} vue principalement à partir de la préface de l'auteur et deux notices bibliographiques rédigées par des lettrés illustres," *Historia Mathematica*, **20**, 160–179.
- Martzloff, Jean-Claude, 1994. "Chinese mathematics," in: I. Grattan-Guinness, ed., *Companion Encyclopedia of the History and Philosophy of the Mathematical Sciences*, Vol. 1, Routledge, London, pp. 93–103.
- Matvievskaya, G. P., 1999. "On the Arabic commentaries to the tenth book of Euclid's *Elements*," *Istoriko-Matematicheskie Issledovaniya*, **4** (39), 12–25 (Russian).

- Melville, Duncan, 2002. "Weighing stones in ancient Mesopotamia," *Historia Mathematica*, **29**, No. 1, 1–12.
- Menninger, Karl, 1969. *Number Words and Number Symbols: A Cultural History of Numbers*, translated from the revised German edition by Paul Broneer. MIT Press, Cambridge, MA.
- Mikami Yoshio, 1913. *The Development of Mathematics in China and Japan*. Reprint: Chelsea, New York, 1961.
- Mikolás, M., 1975. "Some historical aspects of the development of mathematical analysis in Hungary," *Historia Mathematica*, **2**, No. 2, 304–308.
- Milman, Dean; Guizot, M.; Smith, William, 1845. *The History of the Decline and Fall of the Roman Empire by Edward Gibbon*, John D. Morris, Philadelphia.
- Montet, Pierre, 1974. *Everyday Life in Egypt in the Days of Ramesses The Great*, translated from the French by A. R. Maxwell-Hyslop and Margaret S. Drower, Greenwood Press, Westport, CT.
- Moore, Gregory H., 1982. *Zermelo's Axiom of Choice*, Springer-Verlag, New York.
- de Mora-Charles, S., 1992. "Quelques jeux de hazard selon Leibniz," *Historia Mathematica*, **19**, No. 2, 125–157.
- Murata Tamotsu, 1994. "Indigenous Japanese mathematics, *wasan*," in: I. Grattan-Guinness, ed., *Companion Encyclopedia of the History and Philosophy of Mathematical Science*, Vol. 1, Routledge, London, pp. 104–110.
- Narasimhan, Raghavan, 1990. *Bernhard Riemann: Gesammelte Mathematische Werke, Wissenschaftlicher Nachlaß und Nachträge*, Springer-Verlag, Berlin.
- Needham, J., 1959. *Science and Civilisation in China*, Vol. 3: *Mathematics and the Sciences of the Heavens and the Earth*, Cambridge University Press, London.
- Neeley, Kathryn A., 2001. *Mary Somerville: Science, Illumination, and the Female Mind*, Cambridge University Press, New York.
- Neugebauer, O., 1935. *Mathematische Keilschrifttexte*, Springer-Verlag, Berlin.
- Neugebauer, O., 1952. *The Exact Sciences in Antiquity*, Princeton University Press, Princeton, NJ.
- Neugebauer, O., 1975. *A History of Ancient Mathematical Astronomy* (three vols.), Springer-Verlag Berlin.
- Novosyolov, M. M., 2000. "On the history of the debate over intuitionistic logic," *Istoriko-Matematicheskie Issledovaniya*, **5** (40), 272–280 (Russian).
- Ore, Oystein 1957. *Niels Henrik Abel, Mathematician Extraordinary*. Reprint: Chelsea, New York, 1974.
- Özdural, Alpay, 2000. "Mathematics and arts: connections between theory and practice in the Medieval Islamic world," *Historia Mathematica*, **27**, 171–200.
- Panteki, M., 1987. "William Wallace and the introduction of Continental calculus to Britain: a letter to George Peacock," *Historia Mathematica*, **14**, No. 2, 119–132.
- Parshall, Karen Hunger, 1985. "Joseph H. M. Wedderburn and the structure theory of algebras," *Archive for History of Exact Sciences*, **32**, No. 3/4, 223–349.
- Parshall, Karen Hunger, 1988. "The art of algebra from al-Khwarizmi to Viète: a study in the natural selection of ideas," *History of Science*, **26**, 129–164.
- Parshall, Karen Hunger, 2000. "Perspectives on American mathematics," *Bulletin of the American Mathematical Society*, **37**, No. 4, 381–405.

- Patterson, S. J., 1990. "Eisenstein and the quintic equation," *Historia Mathematica*, **17**, 132–140.
- Pavlov, Ivan, 1928. *Conditioned Reflexes*. Reprint: Dover, New York, 1960.
- Pavlov, Ivan, 1955. *Selected Works*, Foreign Languages Publishing House, Moscow.
- Perminov, V. Ya., 1997. "On the nature of deductive reasoning in the pre-Greek era of the development of mathematics," *Istoriko-Matematicheskie Issledovaniya* **2** (37), 180–200 (Russian).
- Pesic, Peter, 2003. *Abel's Proof*, MIT Press, Cambridge, MA.
- Phili, Ch., 1997. "Sur le développement des mathématiques en Grèce durant la période 1850–1950. Les fondateurs," *Istoriko-Matematicheskie Issledovaniya*, 2nd ser., special issue on mathematical schools.
- Piaget, Jean, 1952. *The Child's Conception of Number*, Humanities Press, New York.
- Piaget, Jean; Inhelder, Bärbel, 1967. *The Child's Conception of Space*, Routledge & Kegan Paul, London.
- Picard, É., ed., 1897. *Œuvres Mathématiques d'Évariste Galois*, Gauthier-Villars, Paris.
- Pitcher, Everett, 1988. "The growth of the American Mathematical Society," *Notices of the American Mathematical Society*, **35**, No. 6, 781–782.
- Poisson, S.-D., 1818. "Remarques sur les rapports qui existent entre la propagation des ondes à la surface de l'eau, et leur propagation dans une plaque élastique," *Bulletin des sciences, par la Société Philomatique de Paris*, 9799.
- Porter, Theodore M., 1986. *The Rise of Statistical Thinking*, Princeton University Press, Princeton, NJ.
- Price, D. J., 1964. "The Babylonian 'Pythagorean triangle'," *Centaurus*, **10**, 210–231.
- Pringsheim, A., 1910. "Über neue Gültigkeitsbedingungen für die Fouriersche Integralformel," *Mathematische Annalen*, **68**, 367408.
- Rajagopal, P., 1993. "Infinite series in south Indian mathematics, 1400–1600," *Proceedings of the CSHPM/SCHPM 19th Annual Meeting*, Carleton University, Ottawa, Ontario, pp. 86–118.
- Rashed, Roshdi, 1989. "Ibn al-Haytham et les nombres parfaits," *Historia Mathematica*, **16**, No. 4, 343–352.
- Rashed, Roshdi, 1990. "A pioneer in anaclastics: Ibn Sahl on burning mirrors and lenses," *Isis*, **81**, No. 308, 464–491.
- Rashed, Roshdi, 1993. *Les mathématiques infinitésimales du IXe au XIe siècle*, Vol. II, Al-Furqan Islam Heritage Foundation, London.
- Rashid, Rushdī, 1994. *The Development of Arabic Mathematics: between Arithmetic and Algebra*, translated by Angela Armstrong, Kluwer Academic, Dordrecht and Boston.
- Reich, Karin, 1977. *Carl Friedrich Gauss: 1777/1977*, Inter Nationes, Bonn–Bad Godesberg.
- van Renteln, M., 1987. *Aspekte zur Geschichte der Analysis im 20. Jahrhundert, von Hilbert bis J. v. Neumann*, Lecture notes, University of Karlsruhe.

- van Renteln; M., 1989. *Geschichte der Analysis im 19. Jahrhundert, von Cauchy bis Cantor*, Lecture notes, University of Karlsruhe.
- van Renteln, M., 1991. *Geschichte der Analysis im 18. Jahrhundert, von Euler bis Laplace*, Lecture notes, University of Karlsruhe.
- Richards, Joan, 1987. "Augustus de Morgan and the history of mathematics," *Isis*, **78**, No. 291, 7–30.
- Robins, Gay; Shute, Charles, 1987. *The Rhind Mathematical Papyrus: An Ancient Egyptian Text*, British Museum Publications, London.
- Robson, Eleanor, 1995. *Old Babylonian coefficient lists and the wider context of mathematics in ancient Mesopotamia 2100–1600 BC*. Dissertation, Oxford University.
- Robson, Eleanor, 1999. *Mesopotamian Mathematics, 2100–1600 BC: Technical Constants in Bureaucracy and Education*, Clarendon Press, Oxford and Oxford University Press, New York.
- Robson, Eleanor, 2001. "Neither Sherlock Holmes nor Babylon: a reassessment of Plimpton 322," *Historia Mathematica*, **28**, No. 3, 167–206.
- Rosen, Frederic, 1831. *The Algebra of Mohammed ben Musa*, Oriental Translation Fund, London.
- Rota, Gian-Carlo, 1989. "Fine Hall in its golden age: remembrances of Princeton in the early fifties," in: Duren, 1989, Vol. 3, pp. 223–236.
- Rowe, David, 1997. "Research schools in the United States," *Istoriko-Matematicheskie Issledovaniya*, special issue, 103–127 (Russian).
- Russell, Alex Jamieson, 1879. *On Champlain's Astrolabe, Lost on the 7th June, 1613, and Found in August, 1867*, Burland-Desbarats, Montreal.
- Russell, Bertrand, 1945. *A History of Western Philosophy*, Simon and Schuster, New York.
- Sabra, A. I., 1969. "Simplicius's proof of Euclid's parallel postulate," *Journal of the Warburg and Courtauld Institute*, **32**, 1–24.
- Sabra, A. I., 1998. "One ibn al-Haytham or two? An exercise in reading the bibliographical sources," *Zeitschrift für Geschichte der Arabisch-Islamischen Wissenschaft*, **12**, 1–50.
- Sanderson, Marie, 1974. "Mary Somerville: her work in physical geography," *Geographical Review*, **64**, No. 3, 410–420.
- Sarkor, Ramatosh, 1982. "The Bakhshali Manuscript," *Ganita-Bharati* (Bulletin of the Indian Society for the History of Mathematics), **4**, No. 1–2, 50–55.
- Schafer, Alice T., 1991. "Mathematics and women: perspectives and progress," *Notices of the American Mathematical Society*, **38**, No. 7, 735–737.
- Scharlau, W., 1986. *Rudolf Lipschitz, Briefwechsel mit Cantor, Dedekind, Helmholtz, Kronecker, Weierstraß*, Vieweg, Deutsche Mathematiker-Vereinigung, Braunschweig–Wiesbaden.
- Scharlau, Winfried; Opolka, Hans, 1985. *From Fermat to Minkowski: Lectures on the Theory of Numbers and Its Historical Development*, Springer-Verlag, New York.
- Servos, John W., 1986. "Mathematics and the physical sciences in America, 1880–1930," *Isis*, **77**, 611–629.

- Sesiano, Jacques, 1982. *Books IV to VII of Diophantus' Arithmetica in the Arabic Translation Attributed to Qusta ibn Luqa*, Springer-Verlag, New York.
- Shanker, Stuart, 1993. "Turing and the origins of AI," *Proceedings of the 19th Annual Meeting of the CSHPM/SCHPM*, Carleton University, Ottawa, Ontario, pp. 1–36.
- Sharer, Robert J., 1994. *The Ancient Maya*, Stanford University Press, Pasadena, CA.
- Shen Kangshen, 1988. "Mutual-subtraction algorithm and its applications in ancient China," *Historia Mathematica*, **15**, 135–147.
- Siegmund-Schultze, Reinhard, 1988. *Ausgewählte Kapitel aus der Funktionenlehre*, Teubner, Leipzig.
- Siegmund-Schultze, Reinhard, 1997. "The emancipation of mathematical research publication in the United States from German dominance (1878–1945)," *Historia Mathematica*, **24**, 135–166.
- Sigler, L. E., transl., 1987. *Leonardo Pisano Fibonacci: The Book of Squares*, Academic Press, New York.
- Simonov, R. A., 1999. "Recent research on methods of rationalizing the computation of the Slavic Easter calculators (from manuscripts of the 14th through 17th centuries)," *Istoriko-Matematicheskie Issledovaniya*, **3 (38)**, 11–31 (Russian).
- Singh, Parmanand, 1985. "The so-called Fibonacci numbers in ancient and medieval India," *Historia Mathematica*, **12**, No. 3, 229–244.
- Skinner, B. F., 1948. "'Superstition' in the pigeon," *Journal of Experimental Psychology*, **38**, No. 1 (February), 168–172.
- Smith, David Eugene, 1929. *A Source Book in Mathematics*, 2 vols. Reprint: Dover, New York, 1959.
- Smith, David Eugene; Ginsburg, Jekuthiel, 1934. *A History of Mathematics in America before 1900*, Mathematical Association of America/Open Court, Chicago (Carus Mathematical Monograph 5).
- Smith, David Eugene; Ginsburg, Jekuthiel, 1937. *Numbers and Numerals*, National Council of Teachers of Mathematics, Washington, DC.
- Smith, David Eugene; Latham, Marcia L., transl., 1954. *The Geometry of René Descartes*. Reprint: Dover, New York.
- Smith, David Eugene; Mikami Yoshio, 1914. *A History of Japanese Mathematics*, Open Court, Chicago.
- Solomon, Ron, 1995. "On finite simple groups and their classification," *Notices of the American Mathematical Society*, **42**, No. 2, 231–239.
- Spicci, Joan, 2002. *Beyond the Limit: The Dream of Sofya Kovalevskaya*, Tom Doherty Associates, New York.
- Srinivasiengar, C. N., 1967. *The History of Ancient Indian Mathematics*, World Press Private, Calcutta.
- Stanley, Autumn, 1992. "The champion of women inventors," *American Heritage of Invention and Technology*, **8**, No. 1, 22–26.
- Stevin, Simon, 1585. *De Thiende*. German translation by Helmuth Gericke and Kurt Vogel. Akademische Verlag, Frankfurt am Main, 1965.

- Strauss, Walter, 1977. *Albrecht Dürer: The Painter's Manual*, Abaris Books, New York.
- Struik, D. J., 1933. "Outline of a history of differential geometry," *Isis*, **19**, 92–121, **20**, 161–192.
- Struik, D. J., ed., 1986. *A Source Book in Mathematics, 1200–1800*, Princeton University Press, Princeton, NJ.
- Stubhaug, Arild, 2000. *Niels Henrik Abel and his Times: Called Too Soon by Flames Afar*, Springer-Verlag, New York.
- Stubhaug, Arild, 2002. *The Mathematician Sophus Lie: It Was the Audacity of my Thinking*, translated from the Norwegian by Richard Daly, Springer-Verlag, Berlin.
- Swetz, Frank, 1977. *Was Pythagoras Chinese? An Examination of Right Triangle Theory in Ancient China*, The Pennsylvania State University Studies, No. 40. The Pennsylvania State University Press, University Park, PA, and National Council of Teachers of Mathematics, Reston, VA.
- Tattersall, J. J., 1991. "Women and mathematics at Cambridge in the late nineteenth century," *Proceedings of the 17th Annual Meeting of the CSHPM/SCHPM*, Queen's University, Kingston, Ontario, pp. 53–66.
- Tee, Garry J. 1988. "Mathematics in the Pacific basin," *British Journal of the History of Science*, **21**, 401–417.
- Tee, Garry J. 1999. "The first 25 years of the New Zealand Mathematical Society," *New Zealand Mathematical Society Newsletter*, No. 76, 30–35.
- Thomas, Ivor, 1939. *Selections Illustrating the History of Greek Mathematics*, Vol. 1, *Thales to Euclid*, Harvard University Press, Cambridge, MA.
- Thomas, Ivor, 1941. *Selections Illustrating the History of Greek Mathematics*, Vol. 2, *Aristarchus to Pappus*, Harvard University Press, Cambridge, MA.
- Todhunter, Isaac, 1861. *A History of the Calculus of Variations in the Nineteenth Century*. Reprint: Chelsea, New York, 1962.
- Todhunter, Isaac, 1865. *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace*. Reprint: Chelsea, New York, 1949.
- Toomer, G. J., 1984. "Lost Greek mathematical works in Arabic translation," *The Mathematical Intelligencer*, **4**, No. 2, 32–38.
- Tropfke, 1902. *Geschichte der Elementarmathematik*, 4. Auflage, completely revised by Kurt Vogel, Karin Reich, and Helmuth Gericke. Band 1: *Arithmetik und Algebra*. Walter de Gruyter, Berlin and New York, 1980.
- Tsaban, Boaz; Garber, David, 1998. "On the rabbinical approximation of π ," *Historia Mathematica*, **25**, No. 1, 75–84.
- Urton, Gary, 1997. *The Social Life of Numbers: A Quechua Ontology of Numbers and Philosophy of Arithmetic*, University of Texas Press, Austin, TX.
- de la Vallée Poussin, Charles, 1896. "Recherches analytiques sur la théorie des nombres (première partie)," *Annales de la Société des Sciences de Bruxelles*, **I**, 20₂, 183–256.
- de la Vallée Poussin, Charles, 1899. "Sur la fonction $\zeta(s)$ de Riemann et le nombre des nombres premiers inférieurs à une limite donnée," *Mémoires couronnés et autres mémoires publiés par l'Académie Royale des Sciences, des Lettres, et des Beaux Arts de la Belgique*, **59**.

- Varadarajan, V. S., 1983. "Mathematics in and out of Indian universities," *The Mathematical Intelligencer*, 5, No. 1, 38–42.
- Vidyabhusana, Satis Chandra, 1971. *A History of Indian Logic (Ancient, Mediaeval, and Modern Schools)*, Motilal Banarsidass, Delhi.
- Volkov, Alexei, 1997. "Zhao Youqin and his calculation of π ," *Historia Mathematica*, 24, No. 3, 301–331.
- van der Waerden, B. L., 1963. *Science Awakening*, Wiley, New York.
- van der Waerden, B. L., 1975. "On the sources of my book *Moderne Algebra*," *Historia Mathematica*, 2, 31–40.
- van der Waerden, B. L. 1985. *A History of Algebra from al-Khwarizmi to Emmy Noether*, Springer-Verlag, Berlin.
- Wald, Robert M., 1984. *General Relativity*, University of Chicago Press.
- Walford, E., transl., 1853. *The Ecclesiastical History of Socrates, Surnamed Scholasticus. A History of the Church in Seven Books*, H. Bohn, London.
- Wantzel, Laurent, 1845. "Démonstration de l'impossibilité de résoudre toutes les équations algébriques avec des radicaux," *Bulletin des sciences, par la Société Philomathique de Paris*, 5–7.
- Waring, E. 1762. *Miscellanea analytica*, Oxford.
- Waterhouse, William C., 1994. "A counterexample for Germain," *The American Mathematical Monthly*, 101, No. 2, 140–150.
- Weil, André, 1984. *Number Theory: An Approach Through History from Hammurapi to Legendre*, Birkhäuser, Boston.
- Whiteside, T. L., 1967. *The Mathematical Papers of Isaac Newton*, Johnson Reprint Corporation, London.
- Witmer, T. Richard, transl., 1968. *The Great Art, or The Rules of Algebra*, by Girolamo Cardano, MIT Press, Cambridge, MA.
- Woepcke, Franz, 1852. "Notice sur une théorie ajoutée par Thábit ben Korrah à l'arithmétique spéculative des Grècs," *Journal asiatique*, 4, No. 20, 420–429.
- Woodhouse, Robert, 1810. *A History of the Calculus of Variations in the Eighteenth Century*. Reprint: Chelsea, New York, 1964.
- Wright, F. A., 1933. *Select Letters of St. Jerome*, G. P. Putnam's Sons, New York.
- Yoshida Kôzaku, 1980. "Mathematical works of Takakazu Seki," *The Mathematical Intelligencer*, 3, No. 3: Reprint of material written for the centennial of the Japan Academy.
- de Young, Gregg, 1995. "Euclidean geometry in the tradition of Islamic India," *Historia Mathematica*, 22, No. 2, 138–153.
- Zaitsev, E. A., 1999. "The meaning of early medieval geometry: from Euclid and surveyors' manuals to Christian philosophy," *Isis*, 90, 522–553.
- Zaitsev, E. A., 2000. "The Latin versions of Euclid's *Elements* and the hermeneutics of the twelfth century," *Istoriko-Matematicheskie Issledovaniya*, 5 (40), 222–232 (Russian).
- Zaslavsky, Claudia, 1999. *Africa Counts: Number and Pattern in African Cultures*, Lawrence Hill Books, Chicago.
- Zeuthen, H. G., 1903. *Geschichte der Mathematik im 16. und 17. Jahrhundert*, Teubner, Stuttgart. Johnson Reprint Corporation, New York, 1966.

Zharov, V. K., 2001. "On the 'Introduction' to the treatise *Suan Shu Chimeng* of Zhu Shijie," *Istoriko-Matematicheskie Issledovaniya*, 2nd Ser., No. 6 (41), 347–353 (Russian).

Zhmud, Leonid, 1989. "Pythagoras as a mathematician," *Historia Mathematica*, 16, No. 3, 249–268.

Zverkina, G. A., 2000. "The Euclidean algorithm as a computational procedure in ancient mathematics," *Istoriko-Matematicheskie Issledovaniya*, 5 (40), 232–243 (Russian).

Subject Index

- A Mathematician's Apology*, 20
abacus, 51, 146
Abbasid Caliphate, 54
Abel–Poisson convergence, 501
Abelian group, 447
abscissa, 476
absolute, 371
absolute continuity, 505
Academy, 43, 269, 285, 293, 317
Accademia dei Lincei, 385
acre, 115
Acta eruditorum, 440, 476
Acta mathematica, 67, 73, 450
actinium, 532
adding machine, 149
addition, 5
addition formulas, 492
Aegean Sea, 42
African Americans, 65
Agisymba, 325
aha computations, 134
Ahmose Papyrus, 34, 129–135, 145, 154–156,
210, 235, 239, 247, 399
Ainu, 115
air, 271
Akhmim Wooden Tablet, 34, 132
akoustikoí, 271
alchemy, 15
 \aleph_0 , 547
Alexandria, 34, 81, 105, 270, 300, 305, 317
Algebra
Abu-Kamil's, 428
al-Khwarizmi's, 430, 434
Omar Khayyam's, 425–426
algebra, 3, 22, 30, 32, 41, 56, 57, 67, 137,
139, 536
Chinese, 28
fundamental theorem, 537
multilinear, 380
von Neumann, 71
algebraic topology, 194
Algeria, 59
algorithm, 40, 56, 146, 155
Euclidean, 136, 164–165, 174, 176, 182,
184, 200, 250, 310, 460
mutual-subtraction, 184, 250
Alhazen's problem, 335
Almagest, 43, 47, 55, 60, 305, 325, 338
alphabet, 3
alternating group, 459
amblytomê, 277
American Academy of Arts and Letters, 66
American Geographical and Statistical Society, 86
American Journal of Mathematics, 66, 68,
73
American Journal of Science, 66
American Mathematical Society, 63, 69, 73,
88, 105
American Philosophical Society, 63
amicable numbers, 179, 185
AMS Colloquium Lecture, 103
Analyse des infiniment petits, 473
analysis
functional, 507
nonstandard, 470
analysis situs, 385
Analyst, 488
analytic function, 455
analytic geometry, 187, 202, 348
analytic set, 547
analytical engine, 150
Andaman, 115
angle, 258
acute, 272
central, 298
inscribed, 298
obtuse, 272
right, 272
trisection, 55, 274, 279–282, 332, 335, 425,
433, 435, 449
Archimedes', 313
Annali di scienze matematiche e fisiche, 73
antanaíresis, 285
anthypaíresis, 285
anti-nepotism rules, 76
antipodal points, 481
Anuyoga Dwara Sutra, 23
AO 6670, 402, 407
AO 8862, 401
Apollo, 276
application of areas, 272–274, 306

- with defect, 272, 306, 310
- with excess, 272, 306, 310
- Arabic mathematics, 32
- arc, 258
- Archimedes
 - axiom of, 470
 - tomb, 300
- arcsine, 256, 476
- Arctic Circle, 325
- area, 3, 7, 30, 129, 301, 320, 330, 469, 503
- areas, application, 272–274, 306
 - with defect, 272, 306, 310
 - with excess, 272, 306, 310
- arithmetic, 26, 30, 57
 - commercial, 60, 430
 - Pythagorean, 164
- arithmetic operations, 14
- arithmetic progression, 406
- Arithmetica* (Diophantus), 170, 410, 423, 428, 434
- Arithmetica universalis*, 440
- arithmētikē*, 412
- arithmology, 506
- Armour Institute of Technology, 103
- Ars magna*, 61, 431
- Artis analyticae praxis*, 62, 433
- Aryabhatiya*, 24, 126, 175
- Aryan civilization, 21
- Association for Women in Mathematics, 79
- associative operation, 451
- astral geometry, 342
- astrolabe, 55, 335
- astrology, 15, 29
- astronomy, 26, 29, 32, 60, 192, 258, 260, 269
- Aswan Dam, 123
- asymptotes, 306, 307
- Athenian Empire, 276
- Athens, 286, 317
- Attic Nights*, 271, 317
- attribute, 294
- Aurillac, 59
- Ausdehnungslehre*, 380, 453
- Australian Mathematical Society, 72
- Auswahlprinzip*, 549
- automorphic function, 455
- average, 141, 159, 297, 401
- average person, 525
- axes of conics, 305
- axiom, 32
- axiom of choice, 547, 549, 552, 557
- Azerbaijan, 58
- Aztecs, 69
- Baghdad, 34, 54, 57, 199, 332, 338
- Baire category theorem, 507
- Bakshali, 23, 404
- Bakshali Manuscript, 12, 404, 420
- Baltimore, 64, 88
- Banach–Tarski paradox, 551
- Banneker's Almanac*, 64
- Barnard College, 75, 102
- barycentric coordinates, 368
- base, 113
- Bayes' rule, 539
- Bayesian, 528
- BBC, 38
- bees, 322
- begging the question, 294
- Beijing, 28
- Beijing Library, 32
- Belgian Academy of Sciences, 556
- bell-shaped curve, 518, 523, 531
- Bergama, 46
- Berkeley, 530
- Berlin, 89, 188, 193
- Berlin Academy of Sciences, 193
- Berlin Papyrus 6619, 400
- Bernoulli number, 194
- Bernoulli trials, 515–517
- Betti number, 384
- Bhagabati Sutra*, 23, 212
- Bhagavad Gita*, 421
- Bhaskara II, 184
- Bianchi identity, 385
- bias, 530
- Bibliotheca mathematica*, 548
- bicylinder, 250
- Bijapur, 25
- binary operations, 141
- binomial distribution, 523
- binomial series, 501
- binomial theorem, 468, 472, 515
- Black Sea, 87
- Blessed Islands, 326
- block printing, 28
- blueprint, 4
- Bodhayana Sutra*, 175, 257
- Bologna, 82, 98
- Bolzano–Weierstrass theorem, 391
- Bombay (Mumbai), 26
- bone
 - Ishango, 6, 14
 - Veronice, 6
- Book on the Resolution of Doubts*, 336
- Book on Unknown Arcs of a Sphere*, 338
- Borel sets, 547
- Bororo, 113
- Boston Museum of Fine Arts, 34
- brachystochrone, 376, 479
- Brahmagupta's theorem, 262
- Brahman, 24
- Brahmasphutasiddhanta*, 25, 54, 145, 177, 211, 262
- Brescia, 61
- Brhatsamhita*, 175
- Brianchon's theorem, 365, 369, 394
- Briers score, 531

- Britain, 38, 325
- British Commonwealth, 72
- British Museum, 34, 140, 241, 244
- broken bamboo problem, 266
- Brooklyn Museum of Art, 34
- Brougham Bridge, 452
- Brouwer fixed-point theorem, 558
- Bryn Mawr College, 67, 77, 88, 99, 103
- bu*, 249, 405
- Buddhism, 21, 23, 50, 119
- Bulletin of the American Mathematical Society*, 452
- Bulletin of the New York Mathematical Society*, 451
- Burali-Forti's paradox, 553
- Bushoong, 9
- Byzantine Empire, 41, 48, 50
- c*, 547
- Cairo, 34
- Caius College, 520
- calculation, 146
- calculators, 129, 145, 150
- calculus, 22, 187, 191
 - foundational difficulties, 474
 - fundamental theorem, 467, 470
 - integral, 470
 - priority dispute, 473
 - rules, 471
- calculus of residues, 494
- calculus of variations, 469, 472
- Calcutta, 26
- calendar, 29, 32, 60, 122, 129
 - Archimedes on, 301
 - civil, 123
 - Gregorian, 124, 157
 - Hebrew, 124
 - Julian, 123, 157
 - revised, 128
 - Julian day, 124
 - lunar, 122, 123
 - lunisolar, 122, 157
 - Maya, 124
 - Muslim, 54, 124
 - revised Julian, 128
 - solar, 122
- Calendar Round, 125, 128
- California State University, 65
- Cambridge, 71, 520, 526
- Cambridge University, 101, 171, 467
- Cambridge University Press, 96
- Cambridge, Massachusetts, 68
- Canada, 38, 330
- Canadian Federation, 66
- Canadian Journal of Mathematics*, 73
- Canadian Mathematical Society, 73
- Canary Islands, 326
- cancellation law, 454
- Cantor-Bendixson theorem, 547
- Cardano formula, 431, 440
- cardinal number, 4, 187, 393, 412, 546
- cardinality, 546
- Carolingian Renaissance, 329
- Carthage, 299
- categoricity, 537
- cathedral schools, 58
- Cauchy convergence criterion, 502
- Cauchy integral formula, 494, 500
- Cauchy integral theorem, 494
- Cauchy sequence, 502
- Cauchy-Riemann equations, 493
- cause, 544
- Cavalieri's principle, 31, 250, 253, 302, 303, 314, 465-466, 487, 488
- Cavendish Laboratory, 526
- CBC, 38
- celestial equator, 261
- celestial sphere, 261
- center of curvature, 373
- center of gravity, 525
- Central America, 37
- central limit theorem, 518, 523
- Centrobaryca*, 325
- Ceyuan Haijing*, 418
- chambered nautilus, 9
- Chandahsutra*, 23
- Chebyshev's inequality, 523, 532
- chi*, 246, 248
- chi-square distribution, 529
- chi-square test, 87
- Chiliades*, 287
- China, 23, 37, 40, 113, 139, 145, 197, 302, 326, 338, 407, 420, 529
- Chinese algebra, 40
- Chinese remainder theorem, 173, 175, 180
- Ching (Manchu) Dynasty, 28
- ch'onwonsul* (algebra), 418
- chord, 258
- circle, 8, 298, 329, 356, 418
 - measurement, 301
 - osculating, 374
 - quadrature, 74, 257, 264, 274-275, 290, 312, 449
 - Egyptian approximation, 236
- cissoid, 281
- citizenship, 139
- Civil War, American, 66
- Clark University, 67, 102
- closed set, 392
- clothing, 4
- Cnidus, 286
- code breaking, 150
- coin-tossing, 517
- Colour and Colour Theories*, 102
- Columbia University, 159
- combination, 515
- combinatorics, 187, 212-214

- combinatory product, 380
- comet, 521
- compactness, 391
- compass, 335
- Compendium*, 471
- completeness, 537, 552, 555
- completing the square, 423
- complex number, 187
- composite number, 165
- composite ratio, 353
- compound interest, 403
- Comptes rendus*, 550
- computer program, 4
- computers, 139
- computing, 146
- conchoid, 280–282
- conditional probability, 524
- conditioning, 8
- cone, 276
 - frustum, 324
 - volume, 298
- conformal mapping, 376, 380
- Confucianism, 28
- conic section, 46, 62, 304–310, 352, 394, 471
- conical projection, 327
- Conics*, 55, 296, 315, 323, 335, 337
- conjugate points, 481
- connectedness, 390
- connectivity of a surface, 384
- conservation law, 475
- conservation of energy, 385
- consistency, 537, 552, 555
- Constantinople, 43, 303, 317
- constructivism, 547
- continuity, 390
 - pointwise, 391
 - uniform, 391
- continuum, 284, 390
- continuum hypothesis, 547
- convergence
 - Abel–Poisson, 501
- coordinate system, 9
- coordinates
 - barycentric, 368
 - Cartesian, 327, 352
 - homogeneous, 370
 - line, 370
 - point, 370
- Copenhagen, 208
- Copper Eskimo, 111
- Cordoba, 332
- corner (Egyptian square root), 400
- corner condition, 483
- cosecant, 260
- coset, 448
- cosine, 260, 268
- Cosmos*, 86
- cotangent, 260
- countable set, 549
- counting, 3
- counting board, 51, 135, 146, 253, 413
- counting rods, 51, 135, 146, 198
- Courant Institute, 69
- Cours d'analyse*, 498, 502
- Cramér's paradox, 356
- Cramér's rule, 356
- Crelle's Journal*, 73
- Crete, 276
- Crimean War, 87
- cross-ratio, 356
- Croton, 271
- cube, 274, 275, 288
 - doubling, 274–279, 313, 314, 433, 435, 449
- cube root, 414
- cubic equation, 202, 206, 401, 417, 425, 437
- cubit, 235, 239
- cun*, 248
- cuneiform, 12, 13, 16, 35, 159, 265, 270
- currency conversion, 428
- curvature, 346, 469
 - Gaussian, 377, 395
 - geodesic, 379
- curvature of a surface, 375
- curves, homologous, 388
- curvilinear problem, 280
- Cutting Off of a Ratio*, 305
- cybernetics, 67
- cycle, 460
- cyclic quadrilateral, 263, 268
- cycloid, 371, 465, 469
 - area, 466
 - tangent to, 465
- cycloidal pendulum, 372, 373
- Cyclops, 166
- cyclotomic equation, 447
- cylinder, 276, 303–304, 360
 - area, 238
 - volume, 298
- Cyzicus, 288
- Danish Academy of Sciences, 376
- Data*, 45, 55, 296, 323, 337, 429
- De arte combinatoria*, 214, 515
- De configurationibus qualitatuum et motuum*, 331
- De divina proportionione*, 358
- De numeris datis*, 429, 435
- De quadratris arithmetica circuli*, 203
- De ratiociniis in ludo aleæ*, 514
- De revolutionibus*, 338
- De Thiende*, 144
- De triangulis omnimodis*, 60, 331, 338
- decimal system, 143, 144, 146, 197
- declination, 261
- decumanus maximus*, 327
- Dedekind cuts, 206
- Dedómena*, 299

- definition, 32
- Della pictura*, 61
- Delos, 276
- demotic, 116
- density, 291
- Department of Education, 105
- derivative, 463, 469, 473
 - notation for, 469, 470
- derived set, 547
- Desargues' theorem, 323
- descriptive set theory, 393, 546
- determinant, 53, 419
- development of a function, 542
- diameters of conics, 305
- dice, 532
- Dictionary of Philosophy and Psychology*, 102
- Dictionary of Scientific Biography*, 57
- difference engine, 150
- difference equation, 67
- differentiable manifold, 378
- differential, 470, 472
- differential equation, 67, 389, 472, 474
- differential geometry, 343
- diformly difform, 330
- dimension, 324, 353, 546
- Diophantine equation, 163, 175, 178, 188, 410
- Dirichlet character, 193
- Dirichlet series, 193, 194
- Dirichlet's principle, 482
- Discourse on Method*, 352
- discrimination, 530
- dispersion, 520
- Disquisitiones arithmeticae*, 174, 192
- Disquisitiones generales circa superficies curvas*, 376
- distribution, 496
 - binomial, 523
 - chi-square, 529
 - Gaussian, 522
 - normal, 519, 522–524, 529, 531
 - Poisson, 523
- divination, 10, 14, 29, 175, 210, 522
- Divine Comedy*, 48
- divisibility, 168, 173
- division, Egyptian, 131
- Doctrine of Chances*, 517
- dogs, perception of shape, 8
- domain of rationality, 451
- Doric, 277
- double difference, 246, 258
- double square umbrella, 250
- doubling, 5, 129
- doubling the cube, 274–279, 288, 313, 314, 433, 435, 449
- doubly periodic function, 493
- dozen, 114
- dram, 114
- Dresden Codex, 37
- Dublin, 452
- Düren, 193
- e*, 517
 - transcendence, 197
- earth, 271
- Easter, 11, 129, 157
- eccentricity, 512
- eclipse, 32
 - solar, predicted by Thales, 270
- École Normale, 72, 448
- École Polytechnique, 62, 72, 486
- edge law, 388
- Edict of Nantes, 517
- EDVAC, 153
- Egypt, 5, 13, 19, 37, 40–42, 113, 129, 139, 145, 270, 271, 296, 317, 407, 409
- Egyptian Museum, 34
- elasticity, 379, 383
- electromagnetism, 85, 192
- element, 271
- Elements*, 26, 32, 43, 44, 55, 58, 164, 169, 269, 275, 284, 290, 301, 312, 319, 332, 337, 429
- ellipse, 8, 9, 277, 314, 356
 - definition, 306
 - eccentricity, 8
 - string property, 315
- elliptic function, 191, 194
- elliptic geometry, 349
- elliptic integral, 90, 450, 507
- emolumentum*, 520
- Emperor Yu, 172
- Encyclopédie*, 486
- energy, thermal, 525
- engineering, 129
- ENIAC, 150
- entropy, 523, 526
- envelope, 375, 376
- epanthēma*, 409, 429
- equation
 - cubic, 55, 61, 202, 206, 401, 423, 425, 432, 434, 437
 - in two variables, 394
 - resolvent, 432
 - cyclotomic, 447
 - differential, 474, 507
 - Diophantine, 188, 404, 410
 - gravitational, 385
 - heat, 478
 - Laplace's, 498
 - Pell's, 181, 190
 - quadratic, 235, 401, 407, 420, 423, 437, 460
 - applications, 434
 - positive roots, 421

- quartic, 431–432, 437, 442
- quintic, 446, 449
- solution by radicals, 431
- wave, 477, 497, 498, 507
- equations
 - Frenet–Serret, 384
 - Mainardi–Codazzi, 384
- equator, 326
- equidistant curve, 350
- equinox, 261
- ergodic theory, 67
- Essai sur une manière de représenter les quantités imaginaires dans les constructions géométriques*, 209
- estimation, 524
- Euclidean algorithm, 136, 164–165, 174, 176, 182, 184, 200, 250, 310, 460
- Euclides ab omni nœvo vindicatus*, 339
- Euler characteristic, 386
- Euler constant, 204
- Euler ϕ -function, 189
- Euler's equation, 480
- Euler's formula, 388
- Europe, 168, 327
- European algebra, 40
- European mathematics, 41
- even number, 165, 200
- event, probability, 513
- events, 540
- evolute, 371
- evolution, 529
- existence, 537, 540
- expectation, 514
- exponent, 403
 - integer, 430
- exponentials, 472
- exterior-angle principle, 296
- extremal
 - strong, 483
 - weak, 481
- face, 8
- faction, 437
- fairness, 139
- Fakhri*, 412, 428
- falconry, 59
- false position, 399
- fang cheng*, 405
- fathom, 114
- fen*, 246
- Fermat prime, 188
- Fermat's last theorem, 3, 47, 83, 189, 390
 - fourth powers, 187
- Fermat's little theorem, 189, 219
- Fermat's principle, 479
- Fibonacci Quarterly*, 182
- Fibonacci sequence, 181–182, 185, 213
- field, 447, 548
- Fields Medals, 68, 71
- figurate numbers, 167–168
- figures, 8
- finger reckoning, 143
- finite group, 457–458
- fire, 271
- first category, 506
- five-line locus, 323
- flat manifold, 382
- flat surface, 375
- Flos*, 428
- fluent, 469, 473
- fluxion, 468–469, 473, 474, 484
- Fluxions*, 373, 476
- focal property, 308, 314
- foci, 314
- folium of Descartes, 465
- foot, 114
- Form, 44
- Formal Logic*, 537, 540
- formula, well-formed, 557
- four-line locus, 323, 354
- Fourier coefficients, 499
- Fourier integral, 500
- Fourier inversion formula, 500
- Fourier series, 71, 499, 507
- fractional-linear transformation, 365, 455
- fractions, 30
 - common, 139
 - sexagesimal, 140
- France, 4, 323
- Franciscans, 37, 40
- French Revolution, 83
- Frenet–Serret equations, 384
- frequentist, 528, 538, 543
- frustum
 - of a cone, 252, 301, 324
 - of a pyramid, 248
 - volume, 240, 244
- function
 - absolutely continuous, 505
 - analytic, 490
 - automorphic, 455
 - definition, 498
 - development, 542
 - doubly periodic, 493
 - generalized, 496
 - harmonic, 502, 505
 - multivalued, 387, 494
- functional analysis, 507, 551
- fundamental group, 389
- fundamental theorem of algebra, 537
- furlong, 114
- furniture, 4
- Göttingen, 376
- Galois resolvent, 448
- Ganges, 119
- GAR, 241
- Gauss–Bonnet theorem, 379

- Gaussian curvature, 192, 377, 395
- Gaussian distribution, 64, 522
- Gaussian domain, 458
- Gaussian elimination, 405
- Gaussian integers, 193, 458
- Gebilde*, 495
- generalized function, 496
- Geneva, 86
- geodesic, 379, 383
- geodesic curvature, 379
- geodesy, 192
- Geography*, 55, 327
- geography, 325
- Geometria*, 329
- Geometria organica*, 355
- geometria situs*, 385
- Geometriae prima elementa*, 342
- geometric algebra, 285
- geometric progression, 406
- Géométrie*, 202, 352, 360, 464
- geometry, 8, 26, 41
 - analytic, 474
 - Euclidean, 470
 - hyperbolic, 378
 - non-Euclidean, 350, 371, 378
 - solid, 298
- George Szekeres Medal, 72
- German Mathematical Union, 451
- Gesetz der Kanten*, 388
- GIMPS, 167, 195
- Girton College, 88, 93
- Goldbach conjecture, 190, 195
- golden number, 157
- Golden Ratio, 9, 39
- Golden Section, 298, 358
- Good Will Hunting*, 217
- googol, 128
- googolplex, 128
- Göttingen, 72
- Göttingen Royal Society, 383
- Göttinger Nachrichten*, 189
- gou*, 245
- gougu* theorem, 245, 248, 266
- gram, 532
- Grammelogia*, 149
- graph theory, 9, 210
- gravitation, 218
- gravitational equation, 385
- Great Pyramid, 270, 310
- greatest common factor, 164, 169
- greatest common measure, 165
- Greece, 407
- Greek mathematics, 41
- Gregorian calendar, 124, 157
- Grolier Codex, 37
- gross, 114
- group, 447, 451
 - Abelian, 447
 - abstract, 454
 - alternating, 459
 - finite, 457–458
 - fundamental, 389
 - homology, 389
 - locally compact Abelian, 551
 - Monster, 458
 - permutation, 459
 - sporadic, 458
 - symmetric, 459
- group representations, 456–457
- Grundgesetze der Arithmetik*, 552
- Grundlagen der Geometrie*, 348
- gu*, 245
- Gwalior, 199
- Haab*, 122, 125
- Hai Dao Suan Jing*, 31, 51, 246
- Halle, 100
- Hamilton, Ontario, 68
- Hamming codes, 20
- Han Dynasty, 27, 30, 31, 119
 - Western, 29, 405
- Handbook of Political Fallacies*, 536
- Hannover, 376
- Hanoi, 326
- Harappa, 21
- harmonic analysis, 67
- harmonic function, 502, 505
- harmonic series, 475
- harpedonáptai*, 235
- Harran, 56
- Harvard University, 67, 68, 101, 102, 453
- hau* computations, 134
- Hausdorff space, 393
- heat equation, 478
- Hebrew calendar, 124
- Heidelberg, 89, 146
- Heine–Borel theorem, 392
- hekat*, 134, 210, 239
- hemisphere, 42
 - area, 238
- heptagon, 55
- heptagonal number, 183
- Heron's formula, 263, 320
- hexagon, 298
- hexagonal number, 168
- hieratic, 34, 116
- hieroglyphics, 34, 116, 154
- Hijra, 124
- Hilbert basis theorem, 97
- Hindu mathematics, 32
- histogram, 87, 529
- historical ordering, 8
- holes, 8
- homogeneous coordinates, 370
- homologous curves, 388
- homology, 389
- homology group, 389

- honeycomb, 322
- Horner's method, 415
- horocycle, 346
- Horologium oscillatorium*, 371
- horosphere, 343, 346
- Horus, 123, 131
- Horus-eye parts, 131
- l'Hospital's rule, 473
- Howard University, 65
- Huguenots, 517
- Hundred Fowls Problem, 406
- Hungary, 60
- hyperbola, 277, 306
 - rectangular, 425
- hyperbolic function, 341
- hyperbolic geometry, 340, 342, 378
- hyperbolic paraboloid, 395
- I Ching*, 10, 172, 522
- Iceland, 325
- icosahedron, 450
- Idea, 44
- ideal number, 459
- Iliad*, 283
- Ilkhan, 58
- imaginary geometry, 345
- inclusion-exclusion principle, 513
- incommensurables, 202, 284–285, 287, 488
- incompleteness theorems, 552
- independent trials, 513
- India, 12, 37, 40, 58, 113, 143, 145, 197, 246, 407, 463
- Indian Statistical Institute, 26
- Indiana, 236
- induction, 538, 540, 543
 - transfinite, 547
- Indus River, 21
- infinite, 295, 421
- infinite precision, 7
- infinite series, 22
- infinitely divisible, 3
- infinitesimals, 470, 490
- infinity, 421
- inflection point, 374
- inheritance, 145
- Institute for Advanced Study, 100
- Institute for Scientific Computing, 179
- Institutiones calculi*, 474
- Institutions de physique*, 82
- insurance, 522, 543
- integral, 463, 473
 - elliptic, 507
 - Fourier, 500
 - non-elementary, 474
 - notation for, 469, 471
 - Riemann, 502
- integral domain, 451
- integrating factor, 475
- integration, 440
- International Congress of Mathematicians, 29, 68, 98, 195, 457
- Introductio in analysin infinitorum*, 474, 490
- Introduction to Set Theory*, 557
- involute, 371
- Ionia, 42
- Iran, 57
- Iraq, 54, 57, 143
- Ireland, 58
- irrational number, 199, 200
- Ishango, 6
- Ishango Bone, 14
- Isis, 123
- Isis and Osiris*, 237
- isoperimetric inequality, 318, 348
- isoperimetric problem, 322
- isotherms, 379, 382
- Istituzioni analitiche*, 474
- Italian Geographical Society, 86
- Italy, 3, 60, 285, 317
- Jacobi inversion problem, 493, 495
- Jainas, 209, 212
- Jainism, 21–22
- Japan, 3, 32, 420
- Japanese Mathematical Society, 73
- Jena, 552
- Jerusalem, 69
- Jesuits, 28, 32, 40
- Jinkō-ki*, 52, 180, 252
- Jiu Zhang Suanshu*, 30, 31, 33, 35, 51, 145, 156, 164, 249, 251, 266, 405, 413, 434
- Johns Hopkins University, 66, 68, 73, 88, 101
- Jones polynomial, 71
- Josephus problem, 180–181
- Journal de l'École Normale Supérieure*, 73
- Journal de l'École Polytechnique*, 73
- Journal de mathématiques pures et appliquées*, 380
- Journal für die reine und angewandte Mathematik*, 73, 90, 446, 452
- Julian calendar, 123, 157
- Julian day calendar, 124
- Jupiter, 177
- jurisprudence, 543
- Kaballah*, 11
- Kai Fukudai no Hō*, 419
- Kaliningrad, 386
- karat, 114
- kardo maximus*, 327
- Kattigara, 326
- Kazan' Physico-Mathematical Society, 344
- Kelvin temperature, 526
- Ketsugi-shō*, 254
- khar*, 239
- khet*, 235
- Khorasan, 57
- Kievan Rus, 338

- Kingdom of Wei, 31
 knitting, 9
Kokon Sampō-ki, 53, 418
 Königsberg, 60, 189, 483
 bridges of, 9, 386
 Korea, 27, 32, 50, 214
 Kuba, 9
kuttaka, 175–177, 184, 198, 420

L'invention nouvelle en l'algèbre, 437
La recherche, 71
Lady's and Gentleman's Diary, 216
 Lambert quadrilateral, 333
 Laplace transform, 71
 Laplace's equation, 498
 Laplace–Beltrami operator, 385
 Latin square, 210, 216
 latitude, 325
latus rectum, 306, 313
 Laurent series, 494
 law of cosines, 332
 law of large numbers, 513, 516, 523
 law of sines, 331
Laws, 45
Laws of Thought, 540
Le progrès de l'est, 93
 league, 114
 least common multiple, 169
 least squares, 64, 192, 521
 leather roll, 34
 Lebesgue integral, 503
Leçons sur le calcul des fonctions, 490
Leçons sur les fonctions discontinues, 506
Lectiones geometricae, 467
 legacy problems, 423–425, 434
 legislated value of π , 236
 Leipzig, 60
 lemniscate, 450, 491, 492
 length, 3, 7, 12, 30, 503
Let's Make a Deal, 16
 lever, Archimedes on, 301
 Leyden, 352
h, 246
Liber abaci, 181, 210, 428, 430
Liber de ludo, 513
Liber quadratorum, 181, 211, 428
 Libya, 81
 Lie algebra, 455–456
 Lie bracket, 456
 Lie group, 455–456
Life of Pythagoras, 271
 lightning, 13
Lilavati, 25, 211
 line, 233, 271, 284
 line coordinates, 370
 line graph, 87
 linear algebra, 103
 Lisieux, 59
Lives of Eminent Philosophers, 80, 270

Loci, 296
 locus, 323, 352
 four-line, 306–310, 314
 three-line, 306, 314
 two-line, 314
 logarithm, 62, 147–149, 158, 472, 476, 517
 logic, 14, 67
 three-valued, 556
The Logic of Chance, 542
 logicism, 553
logistikē, 412
 London, 149, 268, 529
 London Mathematical Society, 73
 Long Count, 125
 longitude, 325
 lottery, 18
 Louvre, 140, 401
 Lower Saxony, 376
 lunar calendar, 122
 lunes, 275
 quadrature, 312
 lunisolar calendar, 122, 157
Luo-shu, 172
 Lyceum, 269, 293

 Maclaurin series, 476
 Madagascar, 10
 Madrid Codex, 37
 magic square, 172, 210
 Mahler Lectureship, 72
 Mainardi–Codazzi equations, 384
 Malagasy, 10, 11, 15
 Manchester University, 71
 Manchu (Ching) Dynasty, 28, 33
 Manchus, 28
 manifold, 382, 546
 flat, 382
 Maoris, 70
 map, 4, 325
Maple, 154
 mapping, conformal, 376, 380
 Maragheh, 58
 Markov chain, 523
 Marxism, 29
 Massachusetts Institute of Technology, 69
Matematicheskii Sbornik, 73
Mathematica, 154, 250
Mathematical Analysis of Logic, 539
 Mathematical Association of America, 63,
 105, 107
Mathematical Correspondent, 63
 mathematical expectation, 514, 519, 523
 mathematical journals
 American, 73
 Canadian, 73
 French, 73
 German, 73
 Japanese, 53
 Swedish, 73

- mathematical research, 73
- mathematics education, 154
- Mathematics Magazine*, 188
- mathēmatikotí*, 271
- Mathematische Annalen*, 88, 98
- Matlab*, 154
- Mato Grosso, 113
- matrix, 405, 413, 451
 - transition, 524
- maximum-likelihood, 524
- Maxwell's demon, 526
- Maya, 40, 69, 80, 197
- McGill University, 71
- mean and extreme ratio, 297, 358
- mean proportional, 312
- measure, 546
 - Borel, 551
- measure of curvature, 377
- measure zero, 505
- measurement, 3
- Mécanique céleste*, 66, 86
- Mecca, 54
- Medinet Habu, 122
- Melencolia*, 214
- Memorandum for Friends*, 180
- Menelaus' theorem, 362, 393
- Mengenlehre*, 392
- Meno*, 201
- Meróë, 326
- Mersenne prime, 167
- Merton College, 331
- Merton Rule, 331
- Meru Prastara* (Pascal's triangle), 213
- Mesoamerica, 197
- Mesopotamia, 12, 13, 21, 41, 145, 197, 271, 407, 409, 414
- metalanguage, 552, 555
- Method*, 250, 301–304
- method of exhaustion, 290, 302, 465–467, 470
- method of infinite descent, 187, 189, 218
- metric space, 392, 507
- metric, p -adic, 558
- Metrica*, 303, 320
- Mexico, 37
- microgram, 533
- Miletus, 42, 270
- Ming Dynasty, 22, 28
- minimal polynomial, 434
- minimal surface, 376, 480
- minute, 114
- Mirifici logarithmorum canonis descriptio*, 62, 147
- Möbius band, 369, 388
- Möbius transformation, 365, 369, 455
- modus ponens*, 544
- Mogul Empire, 22
- Mohenjo Daro, 21
- Mona Lisa*, 4
- monads, 271
- monasteries, 58
- Mongol Empire, 28
- Mongols, 28, 58
- Monster, 458
- Montreal, 71
- de Morgan's laws, 538
- Moscow, 551
- Moscow Mathematical Society, 73
- Moscow Museum of Fine Arts, 34
- Moscow Papyrus, 238–240, 244, 248
- motion, 334
- Mozambique, 325
- mu*, 405
- multilinear algebra, 380
- multiplication, 146–147
 - Egyptian, 130
- multivalued function, 387, 494
- Mumbai (Bombay), 26
- music, 4
- Muslim algebra, 40
- Muslim calendar, 124
- Muslim mathematics, 41
- mutual-subtraction algorithm, 184, 250
- myriad, 114, 120
- Naples, 86
- National Council of Teachers of Mathematics, 63
- National Medal of Science, 69
- National Physical Laboratory, 71
- National Science Foundation, 105
- National University of Mexico, 69
- Nautical Almanac Office, 67, 68
- Naval Observatory, 68
- Navigator*, 66
- negative number, 135, 187, 197, 434
 - square root of, 62
- Nephthys, 123
- Neptune, 86
- von Neumann algebra, 71
- Neumann Prize, 72
- neûsis*, 281, 313, 425
- New South Wales, 70
- New Testament, 43
- New York Historical Society, 34
- New York Mathematical Society, 73
- New York University, 69, 78
- Newnham College, 93
- Nicomachean Ethics*, 179
- Nile, 34, 128
 - annual flood, 123
- Nine Symposium Books*, 272
- Noetherian ring, 454
- non-Euclidean geometry, 296, 378
- nonagonal number, 183
- nonstandard analysis, 470

- normal distribution, 64, 519, 522–524, 529, 531
 normal subgroup, 449
 North America Act, 66
 North Carolina Central University, 65
North China Herald, 174
 Northwestern University, 76
 notation, 432
 notebooks, Ramanujan's, 27
 nothing, 541
 Nova Scotia, 68
 number, 3
 Avogadro, 532
 Bernoulli, 194
 cardinal, 187, 393, 412, 546
 complex, 187, 490, 498
 composite, 165
 deficient, 166
 Fermat, 188, 557
 figurate, 167–168
 golden, 157
 heptagonal, 183
 hexagonal, 168
 ideal, 459
 imaginary, 490
 interpretation, 208
 irrational, 199, 200, 557
 negative, 187, 197, 430, 434, 490
 nonagonal, 183
 octagonal, 183
 ordinal, 187, 210, 393, 545, 553
 pentagonal, 168
 perfect, 166, 168, 170, 179, 183
 prime, 165, 193, 556
 regular, 196
 rational, 159, 187, 429, 557
 real, 187
 square, 168
 superabundant, 166
 transcendental, 203, 449
 triangular, 168
 number theory, 41
 Pythagorean, 298
 numbers
 amicable, 179, 185
 relatively prime, 165, 169
 numerals
 Hindu–Arabic, 59
 numerology, 15
 Nürnberg, 60

 octagonal number, 183
 odd number, 165, 200
Odyssey, 166
Oedipus the King, 276
oikuménē, 326
On Burning Mirrors, 318
On Exile, 275
On Isoperimetric Figures, 318

On Socrates' Daemon, 288
 one-sided surface, 388
 open set, 392
Optics, 45, 296
 optics, 55
 Archimedes on, 301
 order, 361
 ordering, 546
 ordinal number, 4, 187, 210, 393, 545, 553
 ordinate, 476
 ornamental geometry, 54
orthotomē, 277
 osculating circle, 374
 Osiris, 123
 Ottoman Empire, 50
 outer product, 380
 Oxford University, 101, 331
oxytomē, 277

Pacific Journal of Mathematics, 458
 Pakistan, 21, 25
 Palermo, 384
 palm reading, 15
 Papal States, 82
 Pappus' principle, 249
 Pappus' theorem, 324, 349
 Papyrus
 Ahmose, 34, 156, 239, 247
 Moscow, 238–240
 Reisner, 34
 parabola, 277, 306
 quadrature, 301
 tangent to, 464–465
 paradox
 Banach–Tarski, 551
 Burali-Forti's, 553
 Petersburg, 519
 Russell's, 549, 553, 554, 556
 Zeno's, 283–284, 310
 Achilles, 283
 arrow, 284
 dichotomy, 283
 stadium, 284
paradox curve, 281
 parallel lines, 294
 parallel postulate, 16, 55, 333–338, 350
 parameterized surface, 375
 Paris, 514
 Paris Academy of Sciences, 190, 384, 448, 494
 Paris Codex, 37
Parmenides, 293
 partial fractions, 440
 partial reinforcement, 13
 parts (unit fractions), 129
 table of doubles, 132
 Pascal's theorem, 365
 Pascal's triangle, 213, 515
 Pataliputra, 24

- Patna, 24
- Paulisha Siddhanta*, 24
- pedagogical ordering, 8
- pedagogy, 31
- Pell's equation, 181, 190
- Peloponnesian War, 275, 285
- pendulum, 474
- Pensées*, 364
- pentagon, 9, 298
- pentagonal number, 168
- pentakaidecagon, 298
- perfect number, 168, 170, 179, 183
- perfect set, 547
- Perga, 305
- Pergamum, 46
- Period of Warring States, 27
- permutation, 215, 445, 515
- permutation group, 459
- Persia, 57, 58
- perspective, 85, 356, 358
- Peshawar, 23, 404
- pesu*, 134, 138, 235
- Petersburg, 188
- Petersburg Academy of Sciences, 73
- Petersburg paradox, 519
- Phænomena*, 45, 55, 296
- Philosophical Transactions of the Royal Society*, 525
- philosophy, 329
- Physical Geography*, 86
- Physics*, 283, 290
- physikoi*, 271
- Physique social*, 525
- π , 233, 247, 257, 263, 449
 - "biblical" value, 236
 - irrationality, 341
 - legislated value, 236
 - transcendence, 197, 203
- pie chart, 87
- pigeons, 13
- Piraeus, 288
- Pisa, 73
- place-value system, 113
- planar problem, 280
- plane, 271
- plane trigonometry, 338
- Platonicus*, 276
- Playfair's axiom, 85
- Plimpton 322, 182
- Poincaré conjecture, 390
- point, 284
- point coordinates, 370
- pointwise continuity, 391
- Poisson distribution, 523
- polygon
 - 17-sided, 192, 447, 450
 - circumscribed, 298
 - inscribed, 298
- polyhedron, 271, 388
- polynomial
 - Jones, 71
 - symmetric, 460
- Porisms*, 296
- Portugal, 33
- postulate, 32
- pound, 114
- power series, 474, 486, 489, 503
- Practical Geometry*, 329
- Prague Scientific Society, 446
- Prasum, 325
- predicate, 294
- predicate calculus, 552
- premier* (unknown), 430
- prime, 460
 - Mersenne, 167
- prime decomposition, 180
- prime number, 165, 556
- prime number theorem, 219
- prime numbers, infinitude, 169
- Princeton University, 63, 70, 100, 153, 454
- Principia*, 63, 66, 82, 373, 468, 469, 479
- Principia mathematica*, 554
- Principles of Mathematics*, 553
- Prior Analytics*, 293
- prism, hexagonal, 323
- probability, 67, 210, 472, 503, 532, 538–540
 - conditional, 524
 - of an event, 513
- progression
 - arithmetic, 406
 - geometric, 406
- projection, 62, 385
- proof by contradiction, 556
- proper class, 556
- proportion, 14, 30, 270, 297, 301
 - Eudoxan theory, 298, 324
- proposition, 32
- propositional calculus, 537
- prosthaphæresis, 146–147, 158
- protomathematics, 3
- Prussian Academy of Sciences, 188
- Psammîtēs*, 120
- pseudosphere, 384, 395
- Ptolemais, 81
- Pulkovo Observatory, 344
- Putnam Examination, 107
- pyramid, 42, 301, 310
 - frustum, volume, 248, 263, 264
- Pythagorean comma, 18
- Pythagorean theorem, 31, 111, 159, 237, 238, 242, 245, 248, 257, 271, 297, 310, 343, 349, 400, 418
 - generalizations, 298, 322, 334
- Pythagorean triples, 175, 177, 312
- Pythagoreans, 11, 20, 43, 45, 80, 164, 275, 287, 297, 536

- Qin Dynasty, 27
 quadratic equation, 235, 401, 407, 420, 437, 460
 positive roots, 421
 quadratic formula, 416
 quadratic incommensurables, 297
 quadratic reciprocity, 192, 219
 quadratrix, 275, 280, 281, 310, 353
Quadrature of the Parabola, 324
 quadrilateral, 262
 area, 263, 268
 cyclic, 263, 268
 Lambert, 333
 Saccheri, 333, 350
 Thabit, 333, 338, 339, 350
 quadrivium, 19, 26, 48, 58
 quantum field theory, 71
 quantum mechanics, 67
Quarterly Journal, 452
 quartic equation, 417, 437, 442
 quaternion, 452, 453, 460
 quinquenove, 515
 quintic equation, 446, 449

 Radcliffe College, 102
 radium-228, 532
 radius of curvature, 373
 rainbow, 55
 random variable, 512, 523
 ratio, 291
 composite, 292, 324, 353
 duplicate, 292
 mean and extreme, 297
 rational number, 159, 187, 429
 ratios, first and last, 469
Real and Complex Analysis, 489
 real number, 187
Recherches sur la probabilité des jugemens, 522
 reciprocals, 140
 reciprocity, 459
 rectangle, 235, 333
 rectangular hyperbola, 425
 regular solids, 298, 299, 357
 Archimedes on, 301
 Reisner Papyrus, 34
 relative rate, 469
 relatively prime, 161, 165, 169, 186, 557
 relativity, 67, 383
 Renaissance, 357
 Carolingian, 329
Republic, 19, 286, 312
 residue, 189
 resolvent, 441
 Galois, 448
 retrograde motion, 305
Review of Modern Physics, 38
 revised Julian calendar, 128
 Revolution
 American, 64
 French, 72
Revue scientifique, 93
 Rhind Papyrus, 13, 129–135
 Rhodes, 327
 Riemann hypothesis, 195
 Riemann integral, 502
 Riemann mapping theorem, 482
 Riemann surface, 387, 494
 Riemann zeta function, 194
 Riemann–Roch theorem, 495
 Riemannian manifold, 385
 rigid body motion, 90
rigle des premiers, 430
 ring, 460
 Noetherian, 454
 ring of sets, 393
 Rolle's theorem, 391
 Roman Empire, 317
 Roman numerals, 119, 143
 Rome, 236, 299
 Room squares, 71, 210
 root of an equation, 423, 437
 rope fixers (surveyors), 237
Rough Draft, 362
 row reduction, 405
 Royal Astronomical Society, 86
 Royal Irish Academy, 86
 Royal Society, 66, 468, 473
 RSA codes, 189
Rubaiyat, 57
 Rule of Three, 134, 135, 140, 155, 428
 Russell's paradox, 549, 553, 554, 556
 Russia, 3, 73
 Russian Academy of Sciences, 188
 Russian, number words, 127

 Sabians, 56
 Saccheri quadrilateral, 333, 350
 Sächsische Landesbibliothek, 37
 Saint Andrew's University, 94
 Saint Gerald, 59
 Sakhalin, 115
 Salem, Massachusetts, 65
 Samarkand, 55, 332
 Samos, 42
Sampō Ketsugi-shō, 53
sanbob (abacus), 51
Sand-reckoner, 120
sangaku, 52, 252, 417
 Sanskrit, 21, 23, 112, 118, 212, 233, 330, 422
 Saturn, 176
 Saxony, 37
 scale, musical, 216
 schema, 554
Science, 530
 Scotland, 325
scroll, 360
 Scuola Normale Superiore, 73

- Scythians, 325
 secant, 260, 298, 330
 second, 114
 second category, 506
 second law of thermodynamics, 525
 Second Punic War, 46, 299
seked, 235
 semantics, 537, 554
 semicircle, 275, 329
 semidifference, 141, 159, 297, 401
 semiregular solids, 358
 Senkereh tablets, 140
 sequence, arithmetic, 193
 Sères, 326
 series, 254, 463, 468, 469, 474
 binomial, 501
 Dirichlet, 193
 Fourier, 499, 507
 geometric, 558
 Laurent, 494
 Maclaurin, 474, 531
 power, 474, 486, 489, 503
 Taylor, 472, 496, 501
 trigonometric, 477, 489, 502
 set
 countable, 549
 uncountable, 549
 set theory, 96, 187
 descriptive, 393, 546
 Seth, 123
 Sevastopol, 87
 Seven Years War, 22
 sexagesimal system, 116, 121, 143, 144, 197, 270
 sexual harassment, 79
 Shang Dynasty, 27, 29
 Shang numerals, 135
 Shanghai, 326
 shape, 3
 Sheikh Abd el-Qurna, 234
 Shetland Islands, 325
 Shimura–Taniyama conjecture, 171
 Shogun, 53
 Shogunate Observatory, 73
Shushu Ji, 172
 Sicily, 59, 285, 299
Siddhanta Siromani, 25
 sieve of Eratosthenes, 166
 simply connected surface, 387
 Simpson's paradox, 530, 533
 Sind, 25
 sine, 259, 267, 330
 Sirius, 122
 six-line locus, 323
 slide rule, 148–150
 circular, 149
 slope, 235, 464
 of pyramids, 235
 Snell's law, 335
 Société de Physique et d'Histoire Naturelle, 86
 Society for Industrial and Applied Mathematics, 63
 software, 146
 solar calendar, 122
 solid of revolution, 249
 solid problem, 280
 Song Dynasty, 28
 Söpdit, 122
soroban (abacus), 51
 Sothic cycle, 123
 South Africa, 53
 space, 3
 Hausdorff, 393
 metric, 392
 complete, 507
 topological, 393
 Spain, 4, 33, 43, 54
 sphere, 329, 343, 387
 area, 267, 301–304, 312
 segment, volume, 247
 volume, 249, 253, 257, 301–304
 spherical mapping, 376
 spherical trigonometry, 335
 spiral, 281, 301, 353, 469
 sporadic group, 458
 square, 9
 square number, 168
 square root, 129, 137, 140, 200, 298, 400, 417, 430
 approximation, 141
 squaring the circle, 264, 274–275, 290, 449
 standard deviation, 87, 529
 standard length, 128
 statics, 59
 Statistical Society, 87
 statistics, 14, 515
 Steinmetz solid, 250
Stetigkeit und irrationale Zahlen, 204
Sthananga Sutra, 23
 Stonehenge, 122
 string property, 315
 strong extremal, 483
 Sturm–Liouville problem, 499, 507
Suan Fa Tong Zong, 32, 33, 51, 52
Suan Jing Shishu, 29
suan pan (abacus), 51
Suan Shu Chimeng, 32, 52, 119
 subgroup, 449
 subject, 294
 subtangent, 464, 476
 subtraction, 5
 successor, 552
Suda, 81
Sulva Sutras, 175, 257
 sum of sines, 467

- Summa de arithmetica*, 60, 429
Sun Zi Suan Jing, 51, 135, 137, 145, 156, 164, 172, 210
 Sun, altitude, 246, 261, 266, 268
 surface, 8
 flat, 375
 minimal, 376
 one-sided, 388
 parameterized, 375
 simply connected, 387
 surveying, 30, 129, 233, 246, 260, 271, 327–329
Surya Siddhanta, 23
Sushu Jiu Zhang, 415
 symbol, 3, 410
 symbolic logic, 452
 symmetric group, 459
 symmetric polynomial, 440, 460
Symposium Discourses, 286
Synagōgē, 47, 142, 280, 296, 320, 322, 351
 syntax, 537, 554
Syntaxis, 47, 55
 Syracuse, 46, 285, 299

Taisei Sankyō, 53
 Taiwan, 28
 Tang Dynasty, 28
 tangent, 260, 298, 330, 469
 Fermat's construction, 464–465
 tarot, 15
 Tata Institute, 26
tatamu, 419
 taxation, 157
 taxes, 139
 Taylor series, 476, 496, 501
 Taylor's theorem, 486, 494, 539
 telegraph, 192
tengen jutsu (algebra), 418
 tensor analysis, 385
 tensor product, 380
 tetrahedron, volume, 257
 Thabit quadrilateral, 333, 338, 339, 350
The Analyst, 63
The Connection of the Physical Sciences, 86
The Enfranchisement of Women, 86
The Fable of the Bees, 82
The Subjugation of Women, 86
The Utility of Mathematics, 276
The Woman Inventor, 79
Theatetus, 199
 Thebes, 234
theorem egregium, 378
Théorie analytique de chaleur, 499
Théorie analytique des probabilités, 520
Théorie des fonctions analytiques, 490
Théorie des nombres, 191
Theory of Functions of a Real Variable, 504
 thermal energy, 525
 thermodynamics, 523, 525–527
 second law, 525
 Thoulē, 325
 three-line locus, 323, 354
 thunder, 13
tian yuan (celestial element), 418
Timaeus, 168
 time, 3, 7, 330, 476–477
 measurement, 128
 topology, 8, 15, 546, 551
 torus, 276, 387
 total curvature, 377
 Toulouse, 187, 514
Tractatus de latitudinibus formarum, 330
 tractrix, 343, 347, 395
Transactions of the Royal Irish Academy, 446
 transcendental number, 203, 449
 transfinite, 546
 transfinite induction, 547
 transformation groups, 54
 trapezoid, 235, 266, 329
Trattato d'algebra, 429
Treatise on Optics, 335
Treatise on the Projective Properties of Figures, 367
 Treviso, 144
 triangle, 235, 262, 329
 angle sum, 272
 triangular number, 168
Trigonometriæ, 331
 trigonometric functions, 259, 338, 472
 trigonometric series, 477, 489, 502, 503
 trigonometry, 22, 58, 60, 145, 233, 258–262, 330, 331, 341, 348, 417, 463, 474
 plane, 60, 338
 spherical, 60, 335, 338
 Trinity College, Cambridge, 112
Triparty, 60, 430
 Tripos Examination, 88, 94
 trisection, 55, 274, 279–282, 313, 332, 335, 425, 433, 435, 449
 trochoid, 355
 Troy, 283
 truncated icosahedron, 358
 truth tables, 537
 Tschirnhaus transformation, 442, 450
 Turin, 190
 Turing machine, 152
 Turkey, 56, 270
 two mean proportionals, 276, 286, 313, 314, 433, 435
Tzolkin, 38, 125, 128
 Umayyad Empire, 22, 54
 uncountable set, 549
 undecidable, 552, 559
 unicursal graph, 9, 15
 uniform continuity, 391
 uniform motion, 330

- uniformly diffeomorphic motion, 330
- unique factorization, 200
- unique factorization domain, 458
- unit, 3, 30, 341
- unit price, 138
- United Nations, 28
- United States, 38
- United States Military Academy, 62
- universe, 541
- University College, 529
- University of Adelaide, 70
- University of Amsterdam, 555
- University of Auckland, 70, 71
- University of Berlin, 72
- University of Bologna, 83
- University of Bonn, 369
- University of Breslau, 193
- University of California, 530
- University of Canterbury, 71
- University of Chicago, 67, 103
- University of Erlangen, 97
- University of Göttingen, 77, 94, 97, 192, 386
- University of Helmstedt, 192
- University of Iowa, 102
- University of Jena, 90
- University of Kazan', 344
- University of London, 29, 77, 88
- University of Melbourne, 70
- University of Michigan, 65, 67
- University of Moscow, 73, 504
- University of Otago, 71
- University of Padua, 61
- University of Pavia, 339
- University of Pennsylvania, 100
- University of South Dakota, 102
- University of St. Petersburg, 73
- University of Stockholm, 73
- University of Sydney, 70
- University of Tasmania, 71
- University of Toronto, 68
- University of Virginia, 101
- University of Wisconsin, 300
- Uranus, 86
- urn model, 516, 526, 539
- utility, 532
- Uzbekistan, 55
- Valmiki Ramayana*, 114
- Vandermonde determinant, 444
- vanishing point, 358
- variable, random, 523
- variance, 520
- variety, 389
- Vassar College, 101
- vector, 452
- vector calculus, 453
- Vedas*, 21, 22
- Vega, 344
- velocity, 330, 469
- Venus, 37, 38, 125
- Veronice, 6
- vibrating membrane, 383, 480
- vibrating string, 480, 496–497
- Victoria University, 71
- Vienna, 37, 60, 89
- Viet Nam, 27, 50
- vigesimal system, 121
- Vija Ganita*, 25, 178, 184, 421
- vikalpa*, 212–214
- Vikings, 58
- virtual certainty, 516
- volume, 3, 7, 31, 129, 301
- wall paintings, 4
- War and Peace*, 78
- wasan*, 51, 57, 252, 267
- Washington, D. C., 68
- water, 271
- wave equation, 477, 497, 498, 507
- weak extremal, 481
- weather forecasting, 531
- weaving, 9
- wedge product, 380
- Weierstrass approximation theorem, 505, 506
- Weierstrass *M*-test, 503
- weight, 3, 7
- well ordered, 549
- Wellesley College, 103
- Wesleyan Academy, 101
- West Point, 62
- Western Han Dynasty, 29, 405
- Western Zhou dynasty, 29
- width, 12
- Wilson's theorem, 191
- wolf bone, 6
- Woolwich, 67
- Worcester, Massachusetts, 67
- World War I, 96
- World War II, 96
- XFL Football League, 216
- xian* (bowstring), 245
- Xiangjie Jiuzhang Suan Fa*, 31, 213
- Xugu Suanjing*, 414
- Xugu Zhaiqi Suanfa*, 214
- Yale Babylonian Collection, 36, 243
- Yale University, 65
- yang*, 11
- Yang Hui Suan Fa*, 31
- Yangtze River, 246
- yard, 114
- year, 15
 - sidereal, 127
 - tropical, 127
- Yellow River, 246
- yenri* (circle theory), 254
- ying*, 11

Yuan Dynasty, 28

Zahlkörper, 451

Zahlring, 451

Zaire, 9

zero, 135, 187, 421

 cancellation, 422

zero divisor, 451

ZetaGrid, 195

zetetics, 433

zhang, 248

Zhou Bi Suan Jing, 51, 247, 266, 405

Zhou Dynasty, 27

Zhui Shu, 31

Zürich, 98

Zürich Polytechnikum, 204

Name Index

- Abel, Niels Henrik, 84, 90, 446, 451, 455, 474, 486, 492, 503
Absolon, Karel, 6
Abu Kamil, 434
Abu'l-Wafa, 57, 143
Abu-Kamil, 57, 179, 412, 428
Achilles, 283
Adalbold of Liège, 329
Adrain, Robert, 63
Aesop, 50
Ağargün, Ahmet, 180
Agnesi, Maria Gaetana, 82–83, 87, 474
Akbar the Lion, 22, 26
al-Jayyani, 331
Alberti, Leon Battista, 61, 358
Aleksandrov, Pavel Sergeevich, 547
d'Alembert, Jean le Rond, 441, 477, 486, 496
Alexander Polyhistor, 271
Alexander the Great, 293, 296
Alhazen (ibn al-Haytham), 335, 336
Allman, George Johnston, 242, 280
Althoff, Friedrich, 94
Amenemhet III, 35
Amir-Moez, Ali R., 336, 425
Ampère, André-Marie, 496, 504
Anatolius, 409
Anaxagoras, 269, 275
Ang Tian-Se, 31, 135, 246
Angell, James, 64
Anne (British queen), 525
Antiphon, 290
Apepi I, 35
Apollodorus, 271
Apollonius, 41, 46, 249, 270, 296, 314, 317, 323, 335, 351
Arbuthnott, John, 525, 529
Archimedes, 33, 42, 43, 46, 52, 59, 84, 120, 249, 250, 269, 276, 290, 313, 317, 320, 324, 335, 353
Archytas, 269, 276
Argand, Jean, 209
Aristaeus, 304, 314
Aristarchus, 121
Aristotle, 42, 43, 179, 201, 269, 271, 275, 283, 290, 297, 319, 511
Arjuna, 421
Artin, Emil, 98
Aryabhata I, 45, 118, 126, 198, 267, 420
Ascher, Marcia, 9, 10
Athelhard of Bath, 329
Avogadro, Amadeo, 532
Ayoub, Raymond, 446
Azulai, Abraham, 180

Babbage, Charles, 150
Bagheri, Mohammad, 55
Baigozhina, G. O., 179
Baire, René, 506, 535, 550
Baker, J. N. L., 86
Ball, W. W. Rouse, 94
Baltzer, R., 368
Banach, Stefan, 551
Banneker, Benjamin, 64–65
Baron, George, 63
Barrow, Isaac, 467, 473
Bartels, Johann, 345
Barzin, M., 556
Bashmakova, I. G., 412
Bayes, Thomas, 524
Bell, Eric Temple (John Taine), 452
Beltrami, Eugenio, 346, 384
Belyi, Andrei, 506
Beman, W. W., 283
Bendixson, Ivar, 547
Benedict XIV, 83
Bentham, Jeremy, 536
Berggren, J. L., 55, 335
Berkeley, George, 473, 474, 484, 488
Bernal, Martin, 42
Bernays, Paul, 348, 557
Berndt, Bruce, 27
Bernoulli, Daniel, 477, 496, 519, 532
Bernoulli, Jakob, 472, 517, 519, 523, 524
Bernoulli, Johann, 472, 473, 476, 479, 548
Bernoulli, Niklaus I, 440
Bernoulli, Niklaus II, 519
Bessel, Friedrich Wilhelm, 342, 344
de Bessy, Bernard Frénicle, 189
Betti, Enrico, 194, 384, 449
Bézout, Étienne, 394, 442
Bhaskara II, 45, 198, 211, 428
Bhau Daji, 24

- Bianchi, Luigi, 385
 Bienaymé, Irénée-Jules, 523
 Biggs, N. L., 212, 214
 Birkhoff, George David, 67
 al-Biruni, 24, 57, 320
 Bôcher, Maxime, 103
 Bochner, Salomon, 454
 Boethius, 48, 168, 328
 Boethius, pseudo-, 328
 Bólyai, Farkas, 342
 Bólyai, János, 343, 346, 371
 Bolzano, Bernard, 209, 391, 504
 Bombelli, Rafael, 61, 206, 412
 Boncompagni, Baldassare, 59, 428
 Bonnet, Ossian, 379
 Boole, George, 535, 537, 539, 542, 554
 Borel, Émile, 391, 503, 535, 546, 550
 Bosse, Abraham, 362
 Bottazzini, Umberto, 441, 498, 503, 504
 Bourbaki, Nicolas, 393
 Bouvelles, Charles, 465
 Bouvet, Joachim, 33
 Bowditch, Nathaniel, 64–66, 86
 Boyer, Carl, 463
 Bragg, William Henry, 71
 Bragg, William Lawrence, 71
 Brahe, Tycho, 147
 Brahmagupta, 12, 118, 139, 145, 155, 187, 198, 211, 262, 266, 428
 Brandt, Heinrich, 100
 Brauer, Richard, 457
 Bravais, Auguste, 39
 Bravais, Louis, 39
 Brentjes, Sonja, 180
 de Breteuil, Gabrielle-Émilie le Tonnelier, 82
 Bretschneider, Carl Anton, 242
 Brianchon, Charles, 365
 Briggs, Henry, 149
 Brouwer, Luitzen Egbertus Jan, 555, 558
 Brown, Marjorie Lee, 65
 Bruins, Evert Marie, 402
 Buck, R. C., 163
 Buddha, 21
 Bugaev, Nikolai Vasilevich, 506
 al-Buni, 214
 Burali-Forti, Cesare, 553
 Bürgi, Jobst, 332
 Burington, Richard, 153
 al-Buzjani, Mohammed, 423

 Caesar, Julius, 317
 Campbell, Paul, 103
 Cantor, Georg, 210, 392, 496, 505, 535, 536, 544, 548, 553, 557
 Cantor, Moritz, 180, 237
 Cardano, Girolamo, 61, 399, 439, 513, 523, 531
 Carleson, Lennart, 504
 Carslaw, H. S., 71
 Cartan, Élie, 455
 Cassiodorus, Magnus Aurelius, 328
 Catalan, Eugène, Charles, 179
 Cauchy, Augustin-Louis, 72, 84, 368, 387, 446, 451, 459, 478, 486, 498, 500
 Cavalieri, Bonaventura, 325, 466, 467, 473
 Cavendish, Henry, 218
 Cayley, Arthur, 88, 93, 451
 de Champlain, Samuel, 330
 Chandrasekharan, Komaravolu, 26
 Charlemagne, 58
 Charles II, 467
 Charles V, 37
 Chasles, Michel, 356
 du Châtelet, Marquise, 77, 82
 Chaucer, Geoffrey, 126
 Chebyshev, Pafnutii L'vovich, 196, 219, 523
 Cheng Dawei, 32, 51
 Chevalley, Claude, 455
 Chiang Kai-Shek, 28
 Christianidis, Jean, 172
 Chuquet, Nicolas, 147, 430
 Cicero, 42
 Clagett, Marshall, 300, 331
 Clark, Walter Eugene, 257
 Clavius, Christopher, 32
 Clayton, Peter, 123
 Cleopatra, 317
 Closs, Michael, 37, 80, 111
 Codazzi, Delfino, 384
 Coe, Michael, 37
 Cohen, Paul, 547
 Colebrooke, Henry Thomas, 24, 25, 54, 140, 176, 198, 211, 262, 420
 Coleridge, Samuel Taylor, 452
 Collins, James, 473
 Colson, F. H., 126
 Colson, John, 83
 Columbus, Christopher, 74
 Confucius, 27
 Conon, 300
 Constantine, 317
 Coolidge, Julian Lowell, 345, 357, 365, 371, 375
 Copernicus, 338
 Cossali, Pietro, 422
 Cotes, Roger, 472, 518
 Courant, Richard, 98
 Cox, Elbert, 65
 Coxeter, Harold Scott MacDonald, 68
 Cramér, Gabriel, 356
 Crelle, August Leopold, 73
 Croesus, 270
 Crossley, John N., 422
 Cullen, Christopher, 29, 245
 Cyril, 81
 Cyrus of Novgorod, 157

- al-Daffa, Abu Ali, 143, 199
 Dahan-Dalmédico, Amy, 84
 Dante, 48, 74
 Darboux, Gaston, 498
 Darwin, Charles, 529
 Dauben, Joseph, 69
 David, H. A., 530
 Davis, Philip, 352
 Deakin, Michael, 81
 Dean, Nathaniel, 65
 Dedekind, Richard, 204, 390, 451, 457, 546, 548
 Degen, Ferdinand, 446
 Delamain, Richard, 149
 Demosthenes, 42
 Denjoy, Arnaud, 503
 Desargues, Girard, 62, 323, 360
 Descartes, René, 56, 72, 82, 202, 207, 218, 310, 335, 351, 360, 364, 386, 393, 465, 473, 479, 496, 548
 Detlefsen, Michael, 11, 552
 Dick, Auguste, 77
 Dickson, L. E., 167, 174, 177, 179, 187, 447
 Diels, Hermann, 44
 Dijksterhuis, E. J., 320
 Dilke, O. A. W., 326, 328
 Dinostratus, 275, 280
 Diocles, 318
 Diocletian, 317
 Diogenes Laertius, 43, 80, 270, 271, 280
 Dion, 285
 Dion Cassius, 126
 Dionysus I, 285
 Diophantus, 12, 47, 61, 62, 176, 181, 187, 404, 409, 423, 428, 433
 Dirichlet, Peter, 193, 391, 457, 499, 502
 Donaldson, Simon, 71
 Dorofeeva, A. V., 481
 Dositheus, 300, 301
 Du Bois-Reymond, Paul, 506
 Du Shiran, 31–33, 172, 198, 214, 245, 405, 413
 Du Zhigeng, 32
 Dudley, John, 335
 Dudley, Underwood, 180
 Duillier, Nicolas Fatio de, 473
 Dummit, David, 450
 Dunaij, Cecilia Krieger, 78
 Dunaij, Cypra, 68
 Dupin, Pierre, 375
 Dürer, Albrecht, 214, 356, 373
 von Dyck, Walther, 451
 Dzielska, Maria, 81
 Edward VI, 432
 Edwards, A. W. F., 530
 Edwards, Harold, 459
 Egorov, Dmitrii Fyodorovich, 506
 Ehrman, Esther, 82
 Einstein, Albert, 98, 383, 527
 Eisenstein, Ferdinand, 449
 Ellicott family, 64
 Emperor Yu, 245
 Eratosthenes, 166, 269, 276, 300, 303, 325
 Erdmann, G., 483
 Erdős, Pál, 180
 Errera, A., 556
 Esau, 180
 Escher, Maurits, 68
 Euclid, 12, 16, 26, 32, 34, 43–45, 48, 58, 85, 164, 166, 168, 179, 202, 249, 257, 269, 300, 304, 314, 317, 332–338, 348, 354, 357, 409, 429, 556
 Eudemus, 43, 269, 273, 275, 305
 Eudoxus, 41, 46, 269, 286, 299, 488
 Euler, Leonhard, 9, 83, 167, 183, 188, 191, 203, 208, 256, 377, 383, 474, 477, 489, 490, 496, 517, 542
 Eutocius, 43, 276, 286, 300, 305, 320
 di Fagnano, Giulio de' Toschi, 491
 al-Farisi, Kamal al-Din, 180
 Farr, William, 87
 Farrar, John, 85
 Fawcett, Philippa, 94
 Feigenbaum, L., 476
 Feingold, Mordechai, 473
 Feit, Walter, 29, 458
 Feller, William, 528
 de Fermat, Pierre, 171, 187, 191, 202, 310, 351, 464–466, 473, 514, 515, 557
 Ferrari, Ludovico, 61, 431
 Ferreirós, José, 546
 del Ferro, Scipione, 61
 Fibonacci, 57, 59, 144, 181, 185, 210, 330, 357, 425
 Field, J. V., 364
 Fields, John Charles, 68
 Fior, Antonio Maria, 61
 Fischer, Ernst, 98
 Fletcher, Colin, 180
 Folkerts, Menso, 329
 Fontana, Niccolò (Tartaglia), 61
 de Fontenelle, Bernard Lebouyer, 486
 Förstemann, Ernst, 38
 Fourier, Joseph, 84, 190, 379, 478, 498, 500
 Fowler, David, 285
 Fraenkel, Adolf, 451, 557
 della Francesca, Piero, 357
 Franchella, Miriam, 555
 Franci, Rafaella, 429
 Franklin, Benjamin, 63
 Fraser, Craig, 481, 486
 Fréchet, Maurice, 392
 Frederick I, 59
 Frederick II of Prussia, 188
 Frederick II of Sicily, 59, 181
 Frege, Gottlob, 544, 552

- Frenet, Jean, 384
 Friberg, Jöran, 162
 von Frisch, Karl, 8
 von Fritz, Kurt, 201
 Frobenius, Ferdinand Georg, 457
 Fukagawa Hidetoshi, 52, 252
 Fuson, Karen, 5, 112
- Galileo, 465, 466
 Galois, Évariste, 451, 455
 Galton, Francis, 519, 527
 Garber, David, 236
 Garciadiego, Alejandro, 69
 Gardner, Milo, 35
 al-Gauhari, 333
 Gauss, Carl Friedrich, 52, 83, 174, 192, 194,
 209, 345, 346, 371, 381, 523, 537, 544
 Gellius, Aulus, 271, 317
 Gelon, 300
 Geminus, 305, 320
 Genghis Khan, 22, 28, 58
 George I, 376
 George IV, 376
 Gerbert, 48, 144, 329
 Gerbillon, Jean-François, 33
 Gerdes, Paulus, 237
 Gergonne, Joseph, 368
 Gerhardt, C. I., 476
 Gerling, Christian Ludwig, 342
 Germain, Sophie, 77, 83–85, 106, 379
 Gibbon, Edward, 81
 Gibbs, Josiah Willard, 453
 Gillings, Richard, 132, 199, 237, 240, 400
 Ginsburg, Jekuthiel, 115
 Glashan, J. G., 68
 Glaucou, 19
 Glivenko, Valerii Ivanovich, 556
 Goddard, William, 64
 Gödel, Kurt, 547, 552
 Goetze, Johann Christian, 37
 Goldbach, Christian, 190
 Goldstine, Herman, 153, 478
 Golenishchev, Vladimir Semënovich, 34
 Goodwin, Edwin J., 236
 Gordan, Paul, 97
 Gorenstein, Daniel, 457
 Gould, Carol Grant, 8
 Gould, James L., 8
 Gow, James, 47, 111
 Grabiner, Judith, 487
 Granville, Evelyn Boyd, 65
 Grassmann, Hermann, 453
 Grattan-Guinness, Ivor, 93, 499, 539, 544,
 546, 548, 553, 554
 Gray, J. J., 274, 285, 333, 338, 346, 364
 Green, George, 494
 Greenleaf, Benjamin, 145
 Gregory VII, 59
- Gregory, James, 467, 469
 Grinstein, Louise, 103
 Grosholz, Emily, 475
 Guizal, Brahim, 335
 Guldin, Habakuk Paul, 325
 Guo Shuchun, 30
- Hadamard, Jacques, 197, 550
 Hairetdinova, N. G., 338
 Halayudha, 212
 Hall, G. Stanley, 102
 Halley, Edmund, 305
 Hamilton, William Rowan, 385, 446, 452
 Hamming, R. W., 20, 21, 38
 Hankel, Hermann, 506
 Hardy, G. H., 20, 27, 38
 Harish-Chandra, 26, 455
 Harnack, Axel, 503
 Harriot, Thomas, 479
 Hart, David S., 63
 Hausdorff, Felix, 392
 Hawkins, Thomas, 454
 Hayashi Takao, 175
 ibn al-Haytham, 179, 191, 333, 341, 350, 479
 He Chengtian, 250
 Heath, T. L., 46, 50, 304, 320, 411
 Hector, 283
 Heiberg, Johann Ludwig, 302
 Heine, Eduard, 391
 de Heinzelin de Braucourt, Jean, 6
 Helicon, 288
 Henri IV, 62, 433, 517
 Henrion, Claudia, 104
 Henry, Alan S., 422
 Hensel, Kurt, 558
 Heracleides, 300, 305
 Heraclitus, 44
 Herbart, Johann Friedrich, 367, 381
 Hermes, Johann, 189, 450
 Hermite, Charles, 204, 450
 Herodotus, 234, 270
 Heron, 303, 325, 409
 Herschel, John Frederick, 124
 Hersh, Reuben, 352, 536
 Heytesbury, William, 331
 Hideyoshi, 51
 Hieron II, 299, 300
 Hilbert, David, 77, 103, 195, 348, 451, 548,
 554
 Hill, George William, 67
 Hipparchus, 41, 325
 Hippasus, 201, 271
 Hippasus, 275, 280, 310
 Hippocrates, 275, 276, 312
 Hobson, E. W., 504
 Hogendijk, Jan, 180, 250, 426
 Homann, Frederick, 63, 329
 Homer, 283
 Horner, William, 415

- de l'Hospital, Marquis, 472, 473, 485
 Høytrup, Jens, 140
 Hugh of St. Victor, 329
 Hughes, Barnabas, 213, 429
 Hulegu, 58
 von Humboldt, Alexander, 86
 Hume, David, 543
 Hutton, Charles, 67
 Huygens, Christiaan, 385, 514, 519
 Hypatia, 47, 81, 105, 170
- Iamblichus, 43, 179, 201, 271
 Inhelder, Bärbel, 8
 Innocent III, 59
 Isodorus, 81
 Isomura Kittoku, 53, 254
 Isomura Yoshinori, 53
- Jacob, 180
 Jacobi, Carl Gustav, 90, 168, 194, 385, 450, 481, 492
 James, Portia, 65
 Jami, Catherine, 33
 Jartoux, Pierre, 256
 al-Jayyani, 338
 Jerome, 43
 Jevons, William Stanley, 535, 544
 Jia Xian, 414
 John of Palermo, 181
 Jones, Alexander, 41
 Jones, V. F. R., 71
 Jordan, Camille, 454
 Jordanus Nemorarius, 59, 213, 435
 Julius Caesar, 42
 Justinian, 317
 Jyeshtadeva, 463, 469, 487
- Kang Xi, 33
 Kant, Immanuel, 488
 al-Karaji, 412, 428
 al-Karkhi, 412, 428
 al-Kashi, 28, 55, 143
 Kasir, Daoud, 57, 425
 Kasner, Edward, 128
 Kazdan, Jerry L., 33
 Keith, Natasha, 79
 Keith, Sandra, 79
 Kelvin, Lord (William Thomson), 526
 Kepler, Johannes, 367, 487, 512
 Khafre, 236
 Khayyam, Omar, 202, 336
 Khinchin, Aleksandr Yakovlevich, 556
 al-Khwarizmi, 54, 56, 422, 429, 434
 Killing, Wilhelm, 455
 King, R. Bruce, 450
 Kingsley, Charles, 81
 Kirkman, Thomas, 216
 Kiro, S. N., 344
- Klein, Felix, 93, 94, 282, 342, 343, 347, 366, 369, 378, 381, 385, 446, 450, 451, 453, 459, 493, 548
 Klein, Jacob, 412
 Knapp, Mary, 9
 Kneser, Adolf, 481
 Knorr, Wilbur, 200, 285
 Kobayashi Shōshichi, 54
 Koblitz, Neal, 50
 Koehler, O., 5
 Kovalevskaya, Sof'ya Vasilevna, 77, 83, 89–93, 105, 453, 478
 Kovalevskii, Vladimir Onufrevich, 89
 Kowalewski, Gerhard, 64, 97
 Kreyszig, Erwin, 478
 Krishna, 421
 Kronecker, Leopold, 451, 536, 546, 547
 Kublai Khan, 28
 al-Kuhi, 335
 Kummer, Ernst Eduard, 196, 459, 546
- Ladd-Franklin, Christine, 77, 100–102
 Lagrange, Joseph-Louis, 83, 177, 191, 441, 444, 448, 451, 477, 480, 486, 490, 505, 539
 Lam Lay-Yong, 30, 135, 248
 Lambert, Johann, 333
 Lamé, Gabriel, 195, 379
 de Landa, Diego, 37, 40
 Langevin, Abbé, 68
 Lao-Tzu, 27
 Laplace, Pierre-Simon, 66, 86, 498, 500, 523, 542
 Lasserre, François, 45, 286
 Laugwitz, Detlef, 381, 487
 Laurent, Pierre, 494
 Lebesgue, Henri, 392, 503, 535, 550
 Lebesgue, Victor-Amédée, 167, 183
 Lefschetz, Solomon, 70
 Legendre, Adrien-Marie, 83, 191, 194, 196, 219, 341, 481, 492
 Leibniz, Gottfried, 72, 82, 149, 203, 214, 216, 364, 438, 475, 483, 488, 490, 515, 531, 535, 536, 553
 Leon, 170
 Leonardo of Pisa (Fibonacci), 57, 59, 144, 181, 185, 210, 330, 357, 425, 430
 Levey, Martin, 428
 Levi-Civita, Tullio, 385
 Li Ang, 413
 Li Rui, 434
 Li Shanlan, 33
 Li Yan, 31–33, 172, 198, 214, 245, 405
 Li Ye, 418
 Libbrecht, Ulrich, 416, 418
 Lie, Sophus, 454
 Liebmann, Heinrich, 345
 Lincoln, Abraham, 14
 Lindemann, Ferdinand, 204, 449

- Liouville, Joseph, 379, 499
 Lipschitz, Rudolf, 205
 Listing, Johann Benedict, 385
 Liu Hui, 31, 246, 247, 252
 Lobachevskii, Nikolai Ivanovich, 346, 371
 Loomis, Elias, 33
 Loria, Gino, 93, 96, 106
 Louis Napoleon, 69
 Louis XIV, 33, 82, 517
 Lovelace, Augusta Ada, 153
 Lull, Ramon, 11, 216, 352
 Luzin, Nikolai Nikolaevich, 504, 547, 551

 Mackay, Alan L., 39, 77
 Maclaurin, Colin, 355, 394, 474, 476
 Maddison, Isabel, 94
 Madhava, 463, 487
 Mahavira, 21
 Mahler, Kurt, 72
 Mainardi, Gaspare, 384
 al-Majriti, 180
 Maltby, Margaret Eliza, 94
 al-Mamun, 54, 56, 422
 Mancosu, Paolo, 483
 Mannheim, Jerome, 391
 Mann, Thomas, 500
 al-Mansur, 54, 143
 Marcellus, 299
 Marco Polo, 28
 Marcus Aurelius, 317
 Marinus of Tyre, 325
 Mark Antony, 317
 Markov, Andrei Andreevich, 523
 Martzloff, Jean-Claude, 32, 33, 50
 Mathieu, Émil, 458
 Matsunaga Ryohitsu, 181, 185
 Matvievskaya, G. P., 332
 Maximilian, 69
 Maxwell, James Clerk, 526
 Mayer, Adolf, 456
 Mayer, Tobias, 64
 de' Mazzinghi, Antonio, 429
 McHenry, James, 64
 Melville, Duncan, 36
 Menaechmus, 269, 277
 Mencius, 27
 Mendelssohn, Fanny, 193
 Mendelssohn, Felix, 193
 Mendelssohn, Rebekah, 193
 Menelaus, 280, 357
 Menna, 234
 Menninger, Karl, 111, 115
 Méré, Chevalier de, 513
 Mermin, Norman David, 38
 Mersenne, Marin, 167, 189, 364, 464
 Mihailescu, Predhu, 179
 Mikami Yoshio, 32, 50, 254, 267, 405, 406, 413, 418

 Mill, John Stuart, 86
 Minding, Ferdinand, 343
 Ming Antu, 33
 Minos, 276
 Mitchell, Maria, 101
 Mittag-Leffler, Gösta, 90
 Möbius, August Ferdinand, 365
 de Moivre, Abraham, 523, 531
 Monbu, 51
 Monge, Gaspard, 375
 Monk, J. Donald, 557
 Montet, Pierre, 122
 Moore, Eliakim Hastings, 103, 451
 Moore, Gregory, 549
 de Mora-Charles, S., 515
 Morawetz, Cathleen Synge, 69, 78
 de Morgan, Augustus, 33, 528, 535, 536, 542, 554
 Mōri Kambei, 51
 Mōri Shigeyoshi, 51, 53
 Moschopoulos, Manuel, 214
 Muir, Thomas, 53
 Müller, Johann (Regiomontanus), 60
 Murata Tamotsu, 50, 53, 74, 257, 267

 Nachshon, Rau, 180
 Napier, John, 62, 147, 149
 Napoleon, 73
 Narasimhan, Raghavan, 383
 Needham, J., 212
 Nehru, Jawaharlal, 26
 Nero, 271
 Nesselmann, G. H. F., 411
 Neugebauer, Otto, 36, 44, 239, 241, 244, 401, 409
 Neumann, Bernhard, 72
 Neumann, Hannah, 72
 von Neumann, John, 153
 Newcomb, Simon, 68
 Newton, Isaac, 43, 52, 63, 72, 73, 82, 83, 202, 218, 256, 350, 355, 479, 518
 Nicomachus, 48, 164, 179
 Nicomedes, 280
 Nightingale, Florence, 86–87
 Nilakanta, 463
 Nipsus, M. Iunius, 328
 Noether, Emmy, 77, 97–100, 454
 Noether, Fritz, 97
 Noether, Max, 88, 97

 Octavian, 317
 Oldenburg, Henry, 438, 468
 Omar Khayyam, 57, 425
 Opolka, Hans, 191
 d'Oresme, Nicole, 56, 59, 202, 429, 475
 Orestes, 81
 Osgood, William Fogg, 102
 Ostrogradskii, Mikhail Vasilevich, 494
 Oughtred, William, 149

- Özdural, Alpay, 54
- Pacioli, Luca, 60, 357, 429
- Pamphila, 271
- Panini, 23
- Pappus, 43, 47, 142, 280, 296, 300, 304, 305, 310, 314, 351, 353, 425
- Parmenides, 44
- Parshall, Karen Hunger, 63, 67, 452
- Pascal, Blaise, 72, 149, 213, 364, 467, 473, 514, 515, 531
- Patterson, S. J., 449
- Pavlov, Ivan Petrovich, 8, 13, 522
- Peacock, George, 535
- Peano, Giuseppe, 552
- Pearson, Karl, 517, 529, 530
- Pedoe, D., 52, 252
- Peet, T. E., 239
- Peirce, Benjamin, 453
- Peirce, Charles Sanders, 101, 457, 535, 536
- Pell, Alexander, 103
- Pell, John, 177
- Pepper, Echo Dolores, 454
- Pericles, 275, 276
- Perminov, V. Ya., 13, 16
- Perott, Joseph, 219
- Perron, Oskar, 503
- Pesic, Peter, 399
- Pestalozzi, Johann Heinrich, 367
- Peter I, 188
- Peter, F., 457
- Phili, Christine, 50
- Philolaus, 44, 285
- Piaget, Jean, 8
- Pincherle, Salvatore, 504
- Pingala, 23, 212
- Pitiscus, Bartholomeus, 146, 332
- Pitt, William, 65
- Planudes, Maximus, 50, 172, 214
- Plato, 19, 21, 38, 40, 42, 43, 168, 199, 201, 269, 297, 352, 511
- Playfair, John, 85
- Plutarch, 42, 237, 270, 288, 299, 300, 317
- Poincaré, Henri, 67, 195, 347, 384, 548
- Poisson, Siméon-Denis, 84, 379, 500, 522, 523, 542
- Polybius, 325
- Pompey, 317
- Poncelet, Jean-Victor, 367
- Price, D. J., 161
- Price, Richard, 524
- Prieto, Sotero, 69
- Pringsheim, Alfred, 500
- Proclus, 43, 44, 168, 269, 272, 273, 280, 294, 300, 312, 320
- Prudhomme, Sully (René François Armand), 18
- Psellus, Michael, 409
- Ptolemy (Egyptian ruler), 276
- Ptolemy Euergetes, 305
- Ptolemy Soter, 45, 317
- Ptolemy, Claudius, 41, 43, 47, 54, 58, 60, 74, 258, 270, 294, 305, 325, 335, 338, 341
- Puiseux, Victor, 387, 494
- Pythagoras, 3, 42, 43, 80, 179, 201, 271
- Pytheas, 325
- Qin Jiushao, 415
- Quetelet, Lambert, 525, 542
- ibn-Qurra, Thabit, 56, 179, 180, 185, 333, 341, 349
- Rajagopal, P., 463
- Ramanujan, Srinivasa, 26, 33
- Ramesses III, 122
- Rashed, Roshdi, 335
- R`ashid, Rushd`i, 57
- Recorde, Robert, 144, 432
- Regiomontanus, 60, 331, 338
- Reich, Karen, 521
- Reisner, George Andrew, 34
- Rhind, Alexander Henry, 34
- Riccati, Jacopo, 83
- Ricci, Matteo, 28, 32, 40
- Ricci-Curbastro, Gregorio, 385
- Richards, Joan, 537
- Rickey, V. Fred, 63
- Riemann, Bernhard, 72, 296, 346, 367, 371, 378, 384, 482, 488, 494, 502, 544
- Ries, Adam, 144
- Riesz, Frigyes, 98, 504
- Rittenhouse, David, 63
- Robert of Chester, 427
- de Roberval, Gilles Personne, 465, 466, 473
- Robins, Gay, 237, 240
- Robinson, Abraham, 69, 487
- Robson, Eleanor, 36, 164, 243
- Roch, Gustav, 495
- Rolle, Michel, 473, 485, 486
- Room, Thomas Gerald, 71
- Rosen, Frederic, 425
- Rota, Gian-Carlo, 70
- Rudin, Walter, 489
- Ruffini, Paolo, 444, 451, 486
- Runge, Carl, 450
- Russell, Alex Jamieson, 330
- Russell, Bertrand, 101, 544, 553, 557
- Rutherford, Ernest, 71
- Saccheri, Giovanni, 333
- ibn Sahl, Abu Saad, 335, 479
- Sansei (Seki Kōwa), 53
- Sawaguchi Kazuyuki, 53, 252, 266, 417, 418
- Scaliger, Joseph, 124
- Scaliger, Julius Caesar, 124
- Scharlau, Winfried, 191, 205
- Schlözer, Dorothea, 106
- Schweikart, Ferdinand Karl, 342

- Schwerdtfeger, Hans, 72
 Scott, Charlotte Angas, 88–89, 94, 105
 von Seidel, Philipp Ludwig, 503
 Seki Kōwa, 50, 52, 53, 73, 252, 417
 Senusret I, 34
 Serret, Joseph, 384
 Sesiano, Jacques, 412
 Sesostri, 234
 Shakespeare, 42
 Shannon, Claude, 115
 Shen Kangshen, 174
 Shih Huang Ti, 29, 31
 Shimura Gorō, 54
 Shute, Charles, 237, 240
 Siegmund-Schultze, Reinhard, 495
 Sigler, L. E., 428
 Simonov, R. A., 157
 Simplicius, 43, 275
 Simpson, Edward Hugh, 530
 Singh, Parmanand, 213
 Skinner, Burrhus Frederic, 13, 14, 522
 Smirnova, G. S., 412
 Smith, D. E., 50, 115, 187, 208, 254, 267, 283, 418
 Snell, Willebrod, 335, 479
 Socrates, 19, 44, 285
 Socrates Scholasticus, 81
 Solomon, Ron, 458
 Somerville, Mary, 85–86
 Sopatros, 287
 Sophocles, 276
 Sotion, 288
 Sporos, 280
 Stäckel, Paul, 345
 Stevin, Simon, 144, 432, 437
 Stirling, James, 517
 Story, William Edward, 66, 73, 101
 Strabo, 325
 Struik, Dirk, 189, 364, 438
 Struve, Friedrich Wilhelm (Vasilii Yakovlevich), 344
 Struve, Vasilii Vasil'evich, 238
 Sturm, Charles, 499
 Suiseth, Richard, 331
 Sun Zi, 30
 Suslin, Mikhail Yakovlevich, 547
 Swetz, Frank, 31, 246
 Sylvester II, 59, 329
 Sylvester, James Joseph, 66, 73, 88, 101, 457
 Synesius, 81
 Synge, John, 68
 Szekeres, George, 72
 Taine, John (Eric Temple Bell), 452
 Tait, Peter Guthrie, 526
 Takebe Katahiro, 53
 Takebe Kenkō, 53, 74, 419
 Tannery, Jules, 550
 Tannery, Paul, 409, 413
 Tarski, Alfred, 551
 Tartaglia, 61, 435
 Taurinus, Franz Adolph, 342
 Tausky-Todd, Olga, 100
 Taylor, Brook, 472
 Taylor, Joan, 189
 Taylor, Richard, 171
 Tee, Garry J., 70
 Thales, 42, 280
 Theaetetus, 269, 286
 Theodorus, 269
 Theomedus, 288
 Theon of Alexandria, 47, 296, 320
 Theon of Smyrna, 43, 276, 318
 Thomas, Ivor, 47
 Thompson, John, 458
 Thomson, William (Lord Kelvin), 526
 Thoreau, Henry, 14
 Thureau-Dangin, François, 402
 Thymaridas, 409, 429
 Timur the Lame, 22, 55
 Todhunter, Isaac, 478, 512, 524
 Tolstoy, Leo, 78
 Torday, Emil, 9
 Torricelli, Evangelista, 465
 Tsaban, Boaz, 236
 Tschirnhaus, Ehrenfried Walther, 442
 Turán, Paul, 33
 Turing, Alan, 150
 al-Tusi, Nasir al-Din, 58, 331, 338
 al-Tusi, Sharaf, 435
 Tzetzes, Johannes, 287
 Uhlenbeck, Karen, 76
 Ulugh Beg, 55
 al-Uqlidisi, 143
 Valéry, Paul, 354
 de la Vallée Poussin, Charles, 197
 Varadarajan, V. S., 26
 Venn, John, 520, 547
 Victoria (British Queen), 87
 Vidyabhusana, Satis Chandra, 535
 Viète, François, 62, 351, 433, 435
 da Vinci, Leonardo, 149, 356
 Vinogradov, Ivan Matveevich, 190
 Vitruvius, 42, 300
 Vivanti, Giulio, 548
 Vogel, Kurt, 237
 Voils, D. L., 163
 Volkov, Aleksei, 250
 Volterra, Vito, 503
 van der Waerden, Bartel Leendert, 12, 98, 130, 238, 241, 323, 400
 Walker, Craig Stewart, 330
 Wallis, John, 207, 209, 468
 Wan Pu-son, 50
 Wang Lai, 33

- Wang Lian-tung, 50
Wang Xiaotong, 414, 416
Wantzel, Laurent, 449
Waring, Edward, 191, 444
Watson, G. N., 27
Weber, Ernst Heinrich, 520
Wefelscheid, Heinrich, 189
Weierstrass, Karl, 72, 83, 89, 385, 391, 453,
477, 478, 481, 487, 493, 495, 503, 504,
548
Weil, André, 191, 193
Wessel, Caspar, 208
Weyl, Hermann, 99, 383, 455, 457
Wheeler, Anna Johnson Pell, 102–103, 106
Wheeler, Arthur Leslie, 103
Whitehead, Alfred North, 89
Whiteside, Thomas, 202, 355, 364, 476
Wiener, Norbert, 67
Wiles, Andrew, 171
Wilson, John, 191
Winston, Mary Frances, 94
Witten, Edward, 71
Wollstonecraft, Mary, 86
Woodhouse, Robert, 478, 481
Wylie, Alexander, 174
Xu Guangchi, 32
Yang Hui, 214, 252
Yoshida Koyu, 52, 180, 252
de Young, Gregg, 26, 58
Young, G. Paxton, 68
Young, Grace Chisholm, 78, 93–97
Young, Jeff, 167
Young, William Henry, 78, 93, 392
Yule, George Udny, 530
Zaitsev, E. A., 329
Zeno, 283, 536
Zenodorus, 348
Zermelo, Ernst, 549, 555
Zeuthen, H. G., 432
Zhang Cang, 31
Zhao Shuang, 29, 245
Zharov, V. K., 32
Zhen Luan, 172
Zhu Shijie, 32, 119
Zu Chongzhi, 31, 250, 252, 253
Zu Geng, 31, 250